AD-A056 772 MITRE CORP BEDFORD MASS F/G 17/2
TESTS RESULTS ADVANCED DEVELOPMENT MODELS OF BISS IDENTITY VERI--ETC(U).
JUL 78 M J FOODMAN F19628-77-C-0001
UNCLASSIFIED MTR-3442-VOL-2 ESD-TR-78-150-VOL-2 NL

1 OF 2
ADA
056772

LEVEL II

(18) ESD-TR-78-150-VOL-II

(19) MTR-3442-VOL-II

TESTS RESULTS ADVANCED DEVELOPMENT MODELS OF BISS IDENTITY VERIFICATION EQUIPMENT. VOLUME II, AUTOMATIC SPEAKER VERIFICATION.

(10) MARTIN J. FOODMAN

(11) JULY 1978

(12) 163 p.

Prepared for

DEPUTY FOR SURVEILLANCE AND NAVIGATION SYSTEMS
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
Hanscom Air Force Base, Massachusetts

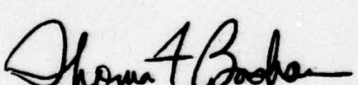
(9) Technical rept.

Project No. 4130
Prepared by
THE MITRE CORPORATION
Bedford, Massachusetts
Contract No. F19628-77-C-0001
(15)

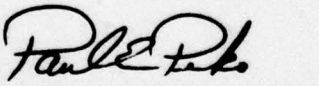
Approved for public release; distribution unlimited.

REVIEW AND APPROVAL

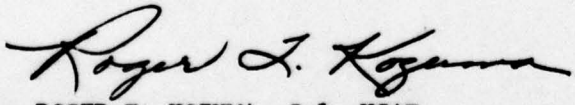
This technical report has been reviewed and is approved for publication.



THOMAS F. BASHARA
Project Manager



PAUL E. PEKO, Lt Col, USAF
Chief, Engineering Division
BIS Systems Program Office



ROGER T. KOZUMA, Col, USAF
Systems Program Director
BIS Systems Program Office
Deputy for Surveillance and Navigation Systems

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ESD-TR-78-150, Vol. II	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Test Results Advanced Development Models of BISS Identity Verification Equipment, Volume II, Automatic Speaker Verification		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER MTR-3442, Vol. II
7. AUTHOR(s) Martin J. Foodman		8. CONTRACT OR GRANT NUMBER(s) F19628-77-C-0001
9. PERFORMING ORGANIZATION NAME AND ADDRESS The MITRE Corporation P.O. Box 208 Bedford, MA 01730		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Project No. 4130
11. CONTROLLING OFFICE NAME AND ADDRESS Deputy for Surveillance and Navigation Systems Electronic Systems Division, AFSC Hanscom Air Force Base, MA 01731		12. REPORT DATE JULY 1978 13. NUMBER OF PAGES 159
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

ACCESS CONTROL
AUTOMATIC SPEAKER VERIFICATION
DIGITAL SIGNAL PROCESSING
ENTRY CONTROL
PERSONAL ATTRIBUTE AUTHENTICATION TECHNIQUES
TESTING
VOICE VERIFICATION

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This volume presents the results of testing the developmental model for speaker personal identity verification. The purpose of the program was to determine the Type I error rate (false rejection of authorized personnel), random Type II error rate (false admittance of unauthorized personnel), and throughput. The tests were conducted both at the Verification Laboratory at The MITRE Corporation and in the

(over)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (continued)

field at the Weapons Storage Area at Pease Air Force Base, New Hampshire. Volume I presents an executive summary of the test results. Volumes III and IV present the detailed results of the handwriting and fingerprint systems, respectively. Volume V presents several miscellaneous but related subjects.

ACKNOWLEDGMENT

This report has been prepared by The MITRE Corporation under Project No. 4130. The contract is sponsored by the Electronic Systems Division, Air Force Systems Command, Hanscom Air Force Base, Massachusetts.

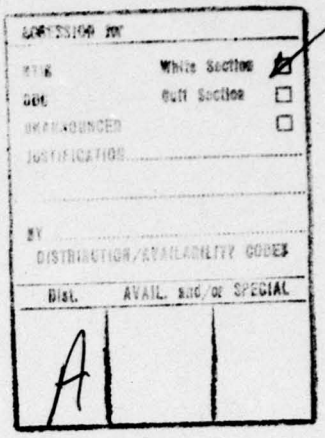


TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

TABLE OF CONTENTS (Cont.)

TABLE OF CONTENTS (Cont.)

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Cont.)

LIST OF TABLES

LIST OF TABLES (Cont.)

1.0 INTRODUCTION

As part of its effort to develop external physical security systems for the Department of Defense, the Electronic Systems Division (ESD) of the United States Air Force, under its Base and Installation Security System (BISS) program, acquired advanced development models of three automated identity verification systems for use in entry control.

The identity verification systems are designed to provide verification of the claimed identity of persons entering or leaving a restricted area through an Entry Control Point. The systems accomplish this verification by examining unique and measurable properties of a person's speech, handwriting and fingerprint. The three systems that were evaluated include the Automatic Speaker Verification (ASV) system, the Automatic Handwriting Verification (AHV) system and the Automatic Fingerprint Verification (AFV) system. The MITRE Corporation was given the responsibility for installing, testing and evaluating the three systems/techniques both at MITRE and at a field test site. This volume, Volume II, of the test report will address itself to the tests, results, and evaluations of the ASV system. Volume I provides an executive summary of the test results and Volume V discusses several miscellaneous but related topics, e.g., human factors tests, hybrid systems.

The basic technique for the ASV system was developed by Texas Instruments Corporation, Dallas, Texas, under contract to the USAF Rome Air Development Center and delivered to MITRE for testing in

July 1975. The ASV system was installed in the Entry Control Laboratory at The MITRE Corporation in Bedford, Massachusetts and tested and evaluated during Phase I and II of the test program. After completion of Phase II, the ASV system was taken to Pease AFB, Portsmouth, New Hampshire for evaluation in a field environment. The ASV system was tested alongside, and in conjunction with the AHV system in all test phases and with the AFV system in Phase II and the Field Test. A detailed description of this system and ASV in general is provided in References 1, 2 and 3.

To use this ASV technique, an individual desiring access into a controlled area arrives at the entry control point and enters his claimed identity, in the form of a four digit identification (ID) number, through a keyboard at the entry terminal. The automatic system then asks the entrant to prove, i.e., to verify, his identity by presenting his credentials. For this system, the credentials are speech material. After appropriate preprocessing, this new material is compared against like material placed in a reference file under that ID at an enrollment session conducted some time earlier. If there is adequate correlation between the two sets of data, the individual is allowed into the controlled area. If not, the individual is given an opportunity to present additional material. When the maximum number of retries is reached and the true owner of the ID number was presenting the credentials and was falsely rejected, a Type I error has occurred. If the person presenting the credentials is not the true owner of the ID and a sufficient correlation existed after any repetition, a false acceptance or a Type II error has occurred. Ideally, neither error should occur. For the

practical systems that are being tested, the Type I and Type II error rates are expected to be less than 1% and 2%, respectively.

The purpose of this test program was to estimate these error rates outside of the contractor's facilities. Other goals were to identify weaknesses and possible improvements to the systems.

1.1 SYSTEM DESCRIPTION

The ASV system has a sensor that detects the entrant's speech and converts it to an analog electrical signal. This signal is then digitized and statistically reduced to a form acceptable for further processing. A keyboard allows the entrant to indicate his claimed identity (i.e., ID number), and an audio indication provides the necessary prompting instructions for the verification sequence. Figure 1 shows a user at an ASV terminal. The Central Processing Unit (CPU) correlates data presented by the entrant with data contained in a reference set obtained during an enrollment process. The reference data is stored on a magnetic disc memory. Additional equipment, which allows an operator to control and monitor the system operation during enrollment and verification, includes a keyboard and such peripheral equipment as a teletype, line printer and card reader for computer input/output operations.

The test program used the CPU for the ASV system as an Integration Processor (IP), which coordinated the operation of the ASV and other verification systems and permitted the data generated by all verification attempts to be recorded on a single digital tape recorder. The combined recorded data was used subsequently to obtain

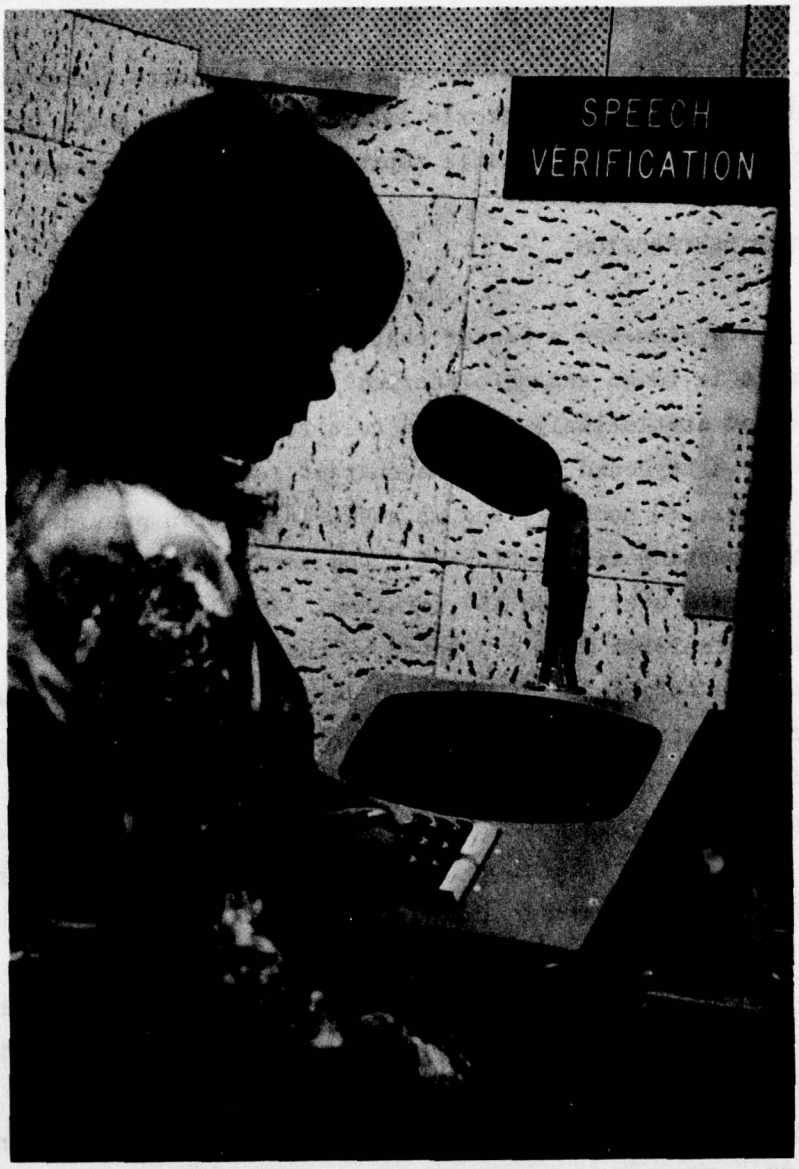


Figure 1 USER AT ASV TERMINAL

the statistical information required concerning error rates and throughput.

1.2 SYSTEM EQUIPMENT

The ASV sensor is a directional dynamic microphone with a frequency response of 50 to 15,000 Hz. It is connected to the preprocessor which converts sensor data from analog to digital. This data then passes through 16 digital filters covering the range of 300 Hz to 3,000 Hz. The output of each filter represents the peak energy in that filter. The ASV software uses only the data from filters 1 through 14.

The entrant terminal is composed of a keyboard and a prompting device. The keyboard consists of the digits 0 through 9, a CLEAR key, and a SEND key. The prompting device consists of a loudspeaker, amplifier, digital/analog converters and CPU interface hardware. Sixteen spoken prompting words, which are combined to form a total of 32 phrases, are stored digitally on a magnetic disc. During enrollment and verification, the CPU pseudo-randomly selects a four-word phrase from the disc, converts it to an analog signal and presents the phrase to the entrant through the loudspeaker at the terminal.

The CPU is a Texas Instruments 980B minicomputer. A disc cartridge memory provides storage space for the reference files, computer programs and the digitized voice library. The latter includes not only the 16 prompting words, but also all digits 0 through 9 and the words or phrases: CALL FOR ASSISTANCE, LOUDER

PLEASE, VERIFIED, NOT VERIFIED, and THANK YOU (in the field test the digits and phrases VERIFIED and NOT VERIFIED were not included in the voice library).

1.3 ENROLLMENT AND VERIFICATION PROCEDURES

During enrollment, the enrollee begins by entering his ID number on the keyboard, and the CPU will cause these digits to be spoken (Phase I and II only) through the prompting loudspeaker as they are entered. If a mistake was made, the enrollee presses the CLEAR button. If satisfied that the correct digits have been entered, the enrollee presses the SEND button. After the operator informs the system that an enrollment is occurring, the CPU selects a four-word phrase at random from the disc and presents it to the enrollee through the loudspeaker. The phrase is 1.9 seconds in duration, and the enrollee is allowed a maximum of 4.0 seconds (a variable under program control) to repeat the phrase. The person must speak into the microphone at a maximum distance of six inches. If a mistake is made while speaking, the enrollee may press the CLEAR button, which causes the system to ignore that repetition of the phrase and to reprompt the same phrase. The enrollee is required to speak a minimum of twenty non-repetitive phrases in a similar manner so that each of the words in the vocabulary are repeated and registered five times. A sufficient number of phrases is processed to construct a reference file of formant frequency structure vs. time for each of the 16 words, where the formants are those frequencies produced by an individual pronouncing particular words and vowel sounds. If during enrollment (or verification) the

words are spoken too softly, the CPU will state through the loudspeaker, "LOUDER PLEASE."

For verification, the individual keys the ID number and presses the SEND button when satisfied that the digits are correct. If the number is not on file, the loudspeaker advises, "CALL FOR ASSISTANCE."

When the CPU receives a valid ID number, it selects, according to a non-repetitive random pattern, a four-word phrase from the disc, and the phrase is presented to the entrant through the loudspeaker. The entrant repeats the phrase as during enrollment. If a mistake is made while speaking, the entrant may press the CLEAR key and the system will ignore that particular phrase. A maximum of one such phrase abort is permitted. For the first four verification attempts immediately following enrollment, referred to as Post Enrollment (PE), the entrant is always required to speak four phrases. Thereafter, the entrant is required to speak only the minimum number of phrases necessary to allow the CPU to make a satisfactory correlation between the reference data and the input data. One four-word phrase is usually sufficient, but as many as eight may be required. When the CPU has completed the correlation, it compares the correlation "score" with the threshold value, and if the score is below the threshold, the entrant is verified. The algorithm used is such that if an individual is not verified on the first phrase, the threshold is adjusted and the subsequent score is based on those obtained for the current and all previous phrases. This process then requires better correlation with the reference data set after the first phrase so that security is maintained. The strategy and the threshold used

during PE are slightly different from the strategy and thresholds used during Normal operation. The test results discussed later show the effects of these differences on error rates.

An individual who claims another's identity will generally be rejected after only four phrases, but in some cases as many as eight phrases are permitted. A Type I error occurs if the scores exceed the threshold for all eight phrases, and the claimed identity is indeed that of the individual seeking verification. The reference data set is adjusted (on a 1/4 basis in PE and on a 1/16 weighted basis thereafter) after each verification to correct for long term voice changes. Serious speech problems such as a very bad cold or laryngitis can cause rejection of a valid verification attempt. The number of people seeking entry under these conditions will be very small, and these speech problems are not expected to degrade the Type I error rate to any significant extent.

1.4 INTEGRATION PROCESSOR

The Integration Processor (IP) coexists with the speaker verification program in the Texas Instruments 980B minicomputer which is part of the ASV system. Basically, the IP managed the operation of the test equipment.

During real-time data collection, the IP had several tasks. During enrollment and verification, it collected data from the ASV or other verification system, placed the data into a standard format

and wrote the data to the line printer and to the digital tape. In addition, for impostor sessions, raw data was written on the magnetic tape for later playback processing. Performance statistics were also maintained and output to the line printer on operator demand.

During playback, the IP replaced the ASV terminals as the source of data for verification. It read raw data from the magnetic tape, sent the data to the appropriate system, received the results and buffered them on disc until it was necessary to write the results onto magnetic tape. As in real time, the IP maintained performance statistics while playback was going on.

2.0 SUMMARY

The test results are summarized in Table I. Table I shows the comparative performance between Phases I and II where the MITRE population was common to both, but the algorithm was different. The table also shows the comparative performance between Phase I and the Field Test where the algorithm was the same, but the populations were different.

Post Enrollment (PE) performance is different from Normal primarily because a different decision strategy was used. The effect of PE on overall performance would be noticeable only during the initial enrollments after installation at a base when a large percentage of the users would be in PE. After that, new enrollees would constitute a small percentage of the total user population.

Table I presents the performance of the male population. Only seven females participated in the Field Test, so no meaningful results of male versus female performance were obtained that could be compared against Phase I. All other objectives presented in the test plans were met. The data collection and analysis revealed that parameters like day of week, time of day, and number of phrases required to enroll did not have a significant effect on ASV performance. Both Type I and Type II error rates in Normal verification were different for females than they were for males, but not with a consistent trend.

Another general result was that errors (Type I and II) were not uniformly distributed among all users -- many users experienced no errors and a few users experienced many errors. This can be

TABLE I

ASV Results for Males

	Phase I		Phase II		Field	
	PE	NOR	PE	NOR	PE	NOR
Number of Users	170	154	164	155	201	131
Type I Error Rate (%)	4.88	0.92	0.16	0.20	9.82	1.09
Type II Error Rate (%)	0.48	0.99	6.43	4.40	0.71	3.26
Verification Time (sec)	18.35	6.54	16.00	5.85	19.98	6.21

PE = Post Enrollment Mode
NOR = Normal Mode

partially explained by looking at the measure of a speaker's consistency, called Expected Scanning Error (ESE). It had a weak influence on the Type I error rate, i.e., users with a high ESE had a somewhat higher incidence of Type I errors. The ESE had a strong influence on the Type II error rate, i.e., the higher the ESE of a user, the higher the probability of a Type II error against that user's reference file.

The BISS requirement of a Type I error rate of 1% was met during the Normal mode of Phase I and Phase II and almost met during the Field Test. The BISS requirement of a Type II error rate of 2% was met during the Normal mode of Phase I but not met during Phase II or the Field Test. Clearly, the Field Test population differed from the MITRE population.

The verification time is the average time from when an individual's identification number is accepted until a decision is made for that person to provide the BISS required throughput using a detention module. The verification time could not exceed six seconds. (See Volume V.) The verification time was almost but not quite achieved by the ASV system.

Limited but more sophisticated Type II testing was conducted against the ASV system using MITRE personnel, college faculty, and drama and speech students. The results are presented in 4.2.4.1.

3.0 OBJECTIVES

The objectives of Phase I, Phase II and the Field Test for the ASV system were the same. The objectives that will be discussed in Sections 4, 5 and 6 are listed below by number, title and paragraph.

No.	Objective	Paragraphs
1	Type I Error Analysis in Real Time	4.2.1,5.2.1,6.2.1
2	Type I Error Analysis in Non-Real Time	4.2.2,5.2.2,6.2.2
3	Independence of Type I Scores	4.2.3,5.2.3,6.2.3
4	Type II Error Analysis in Real Time	4.2.4,5.2.4,6.2.4
5	Type II Error Analysis in Non-Real Time	4.2.5,5.2.5,6.2.5
6	Independence of Type II Scores	4.2.6,5.2.6,6.2.6
7	Sensitivity Analysis of Type I and Type II Errors to Thresholds	4.2.7,5.2.7,6.2.7
8	Verification Time Analysis	4.2.8,5.2.8,6.2.8

A detailed discussion of these objectives including the test approach, data required, data reduction and data analysis is presented in another document. The objectives pertaining to human factors and hybrid systems are discussed in Volume V.

4.0 PHASE I TEST

4.1 DESCRIPTION

The laboratory (Figure 2) used in the Phase I (and Phase II) test was centrally located in MITRE's E building. MITRE personnel were invited to participate in the test activity on a voluntary basis with no particular incentives, rewards or punishment. The data collection period went from 21 October through 26 November 1975. There were two groups of people in the test. The first group of about 50 people participated for the entire six weeks. The second group of about 150 people was enrolled in subgroups of about 50 during the second, third and fourth weeks and each subgroup participated for two weeks.

Participants were enrolled in the test program in groups of four every 45 minutes on Monday and Tuesday. They were then expected to appear at the laboratory at least once per day, but not more than twice per day, morning and afternoon, during their active period. Prior to enrollment, the participants were shown a prerecorded audiovisual briefing using slides and a cassette tape to describe the test program, the operation of the ASV system, and to explain their participation in the program.

A system operator was always present during data collection to initiate enrollment and to observe participant's actions. He also kept a log of those who did not appear at the laboratory on a given day and contacted them by telephone to remind them to stop by the laboratory to verify.

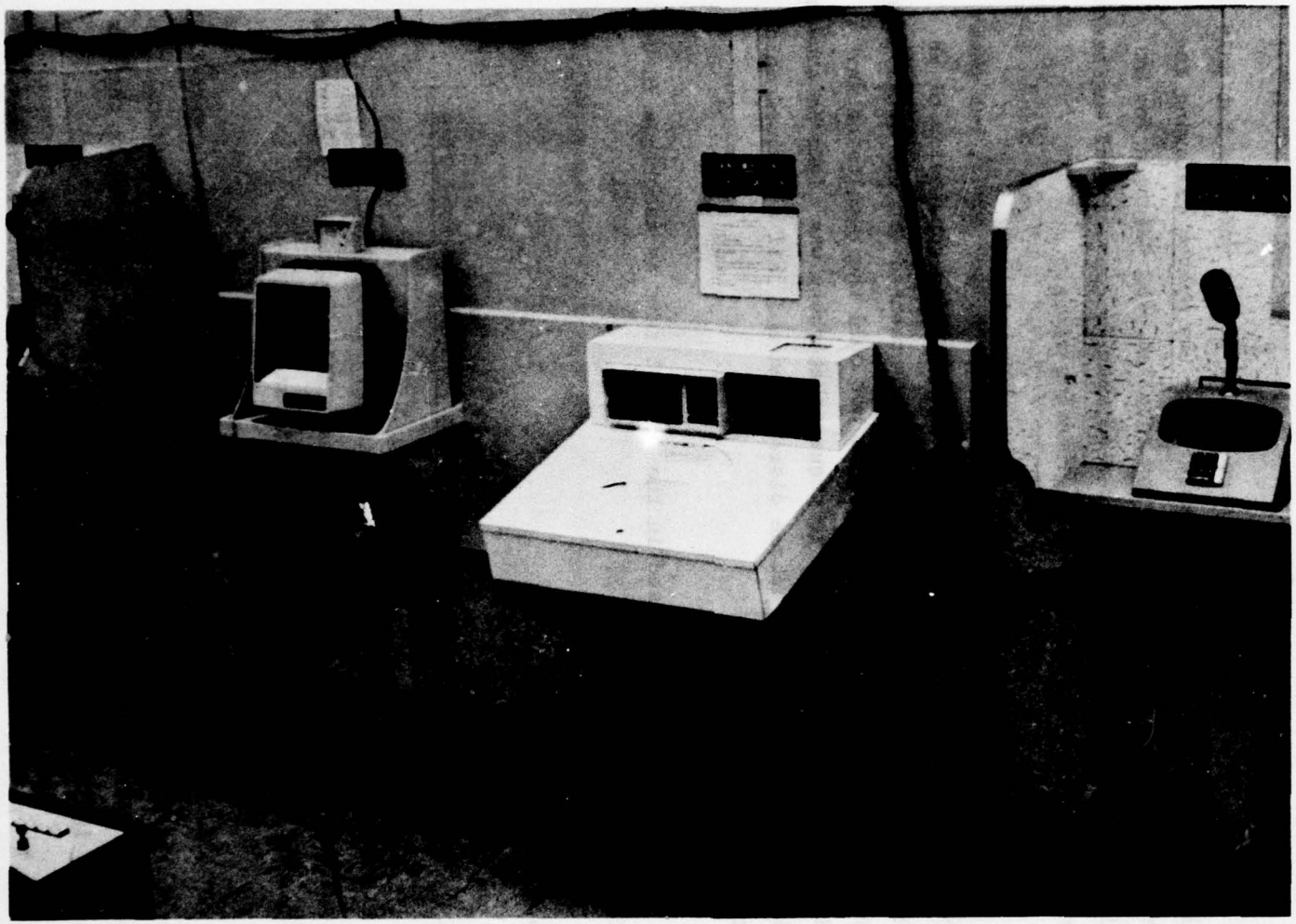


Figure 2 Laboratory for Phases I and II

The test population consisted of 213 people enrolled but only 209 people (170 male and 39 female) used the system following enrollment. Of the total population enrolled, 98% returned for at least one verification. This user population consisted of engineers, secretaries, technicians, and engineering aides.

Participants were asked to do as well as they could, but their performance was not controlled in any way. Upon conclusion of an attempt on the system the participant was thanked (by the loudspeaker on the ASV terminal). Entrants were not told the results of any individual attempt in order to ensure that those who might be causing Type I errors would not become discouraged and fail to return on a regular basis.

The ASV algorithm used in the Phase I test is described in detail in References 1 and 2.

4.2 RESULTS

4.2.1 Objective 1 - Type I Error Analysis in Real Time

To estimate the Type I Error rate (failure to verify proper identity when the representation is actually true) and to determine the confidence limits of the Type I Error estimate.

4.2.1.1 Type I Errors. A Type I error occurs when a person enrolled on the verification system is rejected by the system. Speaker verification Type I errors are grouped according to the sex of the individual.

Individuals with fewer than four successful verifications after enrollment are in Post Enrollment (PE) while individuals with four or more successful verifications are in Normal verification. For example, a person with seven entry verification attempts but only three successful verifications is still in PE. Of the 213 people enrolled in the test, two males and two females were re-enrolled because of their high error rates or high speaker averages. One of the re-enrollees, a male, was re-enrolled a second time. In addition, two other males should have been re-enrolled but were not.

Table II shows the Type I errors, attempts, error rates and upper bounds which occur in both PE and Normal processing. To determine if the PE error rate is significantly different from the Normal verification error rate the F ratio test* is used:

$$T = \frac{1919}{843} \cdot \frac{46}{20} = 5.24$$

$$F_{p=.9}(40,92) = F(40,92) = 1.40$$

Therefore, the error rate in PE is significantly different from the error rate in Normal verification. It appears from this analysis that the PE strategy needs to be modified to reduce the number of

*If the parameter F is less than the test variable T then, at the 90% confidence level, the two test groups (combined males and females in PE and Normal in this case) have significantly different error rates. If F is greater than T then, at the 90% confidence level, it cannot be said that the two test groups have significantly different error rates. See Appendix A.

TABLE II

TYPE I ERROR RATES-ALL USERS

	Errors	Attempts	Error Rate (%)
Post Enrollment	45	843	5.34
Upper Bound*			6.52
Normal	19	1919	0.99
Upper Bound*	—	—	1.35
Total	64	2762	2.32

*90% confident that the true error rate is less than the upper bound. This is the Chi-squared test. See Appendix B.

Type I errors. However, it is also possible that, due to initial nervousness by new users, the PE error rate will always be higher than the Normal strategy. The combined PE and Normal, male and female Type I error rate is 2.32%. This high error rate is due mainly to the very high error rate (5.34%) that occurs during PE. The dominance of PE in the overall performance is due to the short test period.

4.2.2 Objective 2 - Type I Error Analysis in Non-Real Time

To determine how various parameters affect the Type I Errors and system performance.

4.2.2.1 Type I Errors Versus Sex. The first set of ASV error rates shown in Table III are for all errors and all attempts.

TABLE III

TYPE I ERROR RATES VERSUS SEX

Set 1

	Errors	Attempts	Error Rate (%)
Male PE	33	676	4.88
Male Normal	15	1632	0.92
Female PE	12	167	7.19
Female Normal	4	287	1.39
Total	64	2762	2.32

Set 2

	Errors	Attempts	Error Rate (%)
Male PE	18	657	2.74
Male Normal	14	1628	0.86
Female PE	7	154	4.55
Female Normal	3	283	1.06
Total	42	2722	1.54

The error rates in Set 1 include errors occurring before a re-enrollment and include errors of those who should have been re-enrolled. Set 2 error rates exclude errors and attempts of those who should have been re-enrolled but were not, and all errors and attempts before the last re-enrollment for those who were re-enrolled.

The data for Set 2 shows that the males have a lower Type I error rate than females and that the male and female error rates during PE are three and four times, respectively, as high as in Normal verification.

To determine if the male and female error rates are significantly different, the F ratio test is used. Comparing the males and females in PE (Set 2) yields:

$$T = \frac{657}{154} \cdot \frac{8}{19} = 1.79$$

From the F ratio Table (see Reference 8) the corresponding 90% confidence value is:

$$F(38,16) = 1.82$$

Thus, it cannot be said, at the 90% confidence level, that the male and female PE error rates are significantly different.

For Normal verification (Set 2), the test yields:

$$T = \frac{1628}{283} \cdot \frac{4}{15} = 1.53$$

$$F(30,8) = 2.38$$

Once again, it cannot be said that the male and female Normal verification error rates are significantly different.

4.2.2.2 Type I Errors Versus Time Of Day. Type I errors versus time of day is shown in Table IV. The F ratio test for morning and afternoon is:

$$T = \frac{1646}{1027} \cdot \frac{29}{37} = 1.25$$

$$F\ (74,58) = 1.40$$

Therefore, it cannot be said that the error rates between morning and afternoon are statisfically different.

TABLE IV

TYPE I ERROR RATE VERSUS TIME OF DAY

	Errors	Attempts	Error Rate (%)	Upper Bound (%)
Morning	36	1646	2.19	2.73
Afternoon	28	1027	2.73	3.51

4.2.2.3 Type I Error Rates Versus Expected Scanning Error*. The expected scanning error (ESE) is a measure of the consistency between repetitions of the same word. The lower the ESE, the better the match between reference patterns and new speech material. The ESE is used to normalize the current scanning error and determine a decision function which is compared against a threshold. For the decision function computation only, ESE is limited to be in the range between 100 and 140.

Figure 3 shows the Type I error rate versus ESE for males and females for combined Normal and PE data. For this algorithm, the Type I error rate was expected to be higher for those people with an ESE greater than 140. This did not occur.

4.2.2.4 Type I Errors Versus Station. The ASV system had two user terminals called Station 1 and Station 2. Both stations should treat the users the same since the microphones are of the same type, the cable lengths are the same, and the analog and digital circuits are the same.

All enrollments were at Station 1. Verifications could take place at either station, but because of the enrollment experience at Station 1 and the closer proximity of Station 1 to the entrance to the laboratory, most entry attempts took place at Station 1 (see Table V).

*Expected Scanning Error and speaker average refer to the same quantity and are used interchangeably in this report.

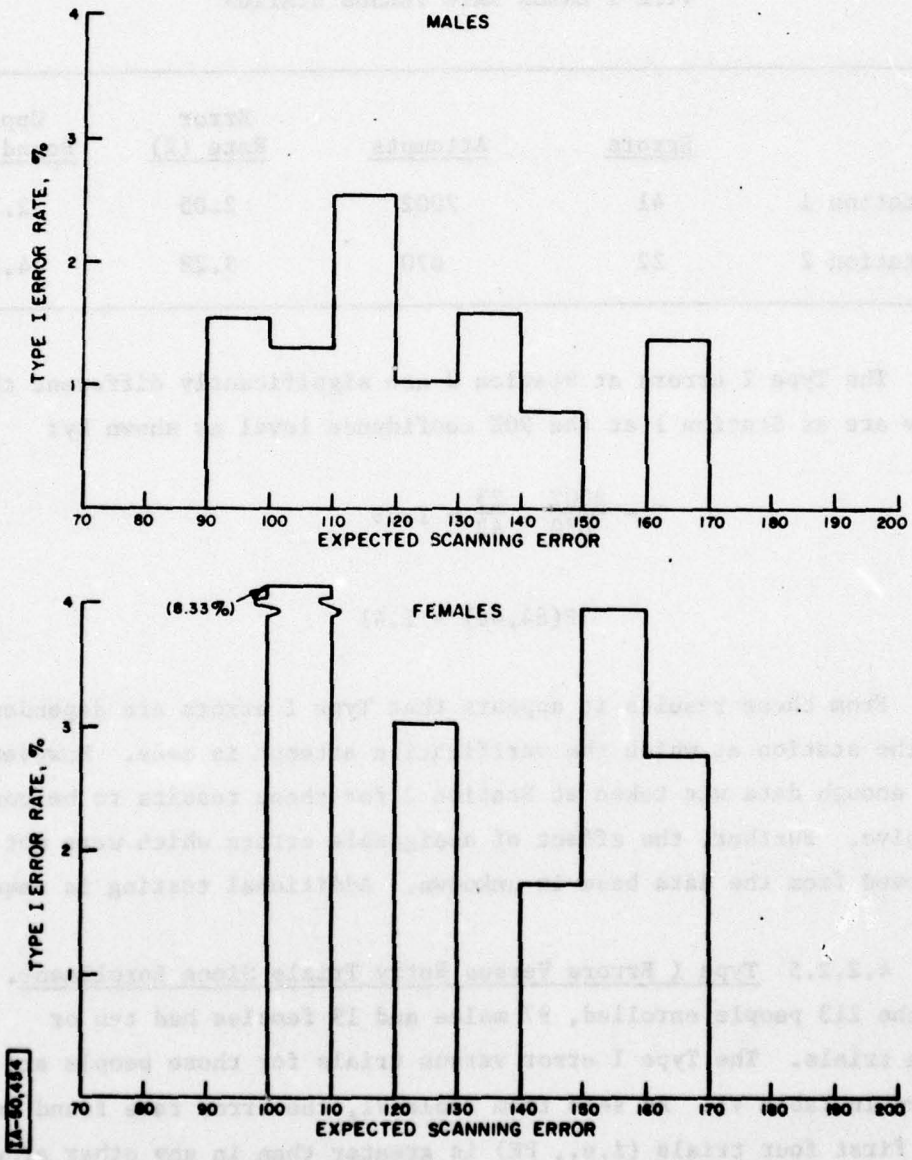


Figure 3 TYPE I ERROR RATE VS EXPECTED SCANNING ERROR

TABLE V

TYPE I ERROR RATE VERSUS STATION

	Errors	Attempts	Error Rate (%)	Upper Bound (%)
Station 1	41	2002	2.05	2.77
Station 2	22	670	3.28	4.38

The Type I errors at Station 2 are significantly different than they are at Station 1 at the 90% confidence level as shown by:

$$T = \frac{2002}{670} \cdot \frac{23}{42} = 1.63$$

$$F(84,46) = 1.41$$

From these results it appears that Type I errors are dependent on the station at which the verification attempt is made. However, not enough data was taken at Station 2 for these results to be conclusive. Further, the effect of assignable errors which were not removed from the data base is unknown. Additional testing is required.

4.2.2.5 Type I Errors Versus Entry Trials Since Enrollment. Of the 213 people enrolled, 97 males and 19 females had ten or more trials. The Type I error versus trials for those people are shown in Table VI. As seen from Table VI, the error rate found in the first four trials (i.e., PE) is greater than in any other group

TABLE VI

TYPE I ERRORS FOR USERS WITH 10 OR MORE TRIALS

Trial Number	Male Errors	Error Rate (%)	Female Errors	Error Rate (%)
1	3	3.09	1	5.26
2	2	2.06	2	10.53
3	2	2.06	2	10.53
4	2	2.00	0	0
5	0	0	0	0
6	1	1.03	0	0
7	0	0	0	0
8	0	0	1	5.26
9	1	1.03	1	5.26
10	1	1.03	0	0

of four trials thereafter. After the first four trials, there are so few Type I errors that no trend in the number of Type I errors versus increasing trial number can be determined. Therefore, more data is required to determine the correlation, if any, between trial number and Type I errors.

4.2.2.6 Type I Errors Versus Phrases Required To Enroll. Type I errors for males and females versus the number of phrases required during enrollment are presented in Tables VII and VIII. The ASV system requires a minimum of twenty phrases to enroll a person. The phrases are prompted until each of the sixteen words is repeated five times. Each of the five repetitions is used to create the reference pattern. If a person says one word in response to a particular prompted word for part of the enrollment, and then says another in response to the same word, five repetitions of the new response are required. The number of phrases, therefore, increases over the minimum of twenty. Other common reasons for added phrases are (1) variable volume during a phrase, (2) variable speed from phrase to phrase, and (3) not completing a phrase due to mike fright, inattention, interruptions, etc. In general, the more phrases required during enrollment, the more difficulty that person had during enrollment. Table VII includes the errors and trials of people who should have been re-enrolled and includes all of the errors and trials of those who were re-enrolled (Set 1). Table VIII excludes the errors and trials of those who should have been re-enrolled and the errors and trials prior to re-enrollment of those who were re-enrolled (Set 2).

TABLE VII

TYPE I ERROR RATE VERSUS PHRASES REQUIRED
DURING ENROLLMENT FOR MALES AND FEMALES - SET 1

	MALE			FEMALE		
Phrases Required	Type I Errors	Entry Attempts	Error Rate (%)	Type I Errors	Entry Attempts	Error Rate (%)
20	20	1214	1.65	7	162	4.32
21	15	530	2.83	5	109	4.59
22	2	71	2.82	2	28	7.14
23	0	85	0.0	1	12	8.33
24	4	38	10.5	0	0	----
25	3	127	2.36	0	38	0.0
26	3	61	4.92	1	54	1.85
27	0	9	0.0	0	0	----
28	0	12	0.0	0	27	0.0
29	0	35	0.0	0	0	----
30	0	0	---	0	0	----
31	0	0	---	0	0	----
32	1	12	8.33	0	0	----
33	0	0	---	0	0	----
34	0	0	---	0	0	----
35	0	0	---	0	0	----
36	0	11	0.0	0	0	----
37	0	0	---	0	0	----
38	0	0	---	0	0	----
≥ 39	0	14	0.0	0	24	0.0
≤ 21	35	1744	2.01	12	271	4.43
≥22	13	475	2.74	4	183	2.19

TABLE VIII

TYPE I ERROR RATE VERSUS PHRASES REQUIRED DURING ENROLLMENT FOR MALES AND FEMALES - SET 2

	MALE			FEMALE		
Phrases Required	Type I Errors	Entry Attempts	Error Rate (%)	Type I Errors	Entry Attempts	Error Rate (%)
20	18	1212	1.49	7	162	4.32
21	4	519	0.77	5	109	4.59
22	2	71	2.82	0	26	0.0
23	0	85	0.0	1	12	8.33
24	3	37	8.11	0	0	0.0
25	3	127	2.36	0	38	0.0
26	0	58	0.0	1	54	1.85
27	0	9	0.0	0	0	----
28	0	12	0.0	0	27	0.0
29	0	35	0.0	0	0	----
30	0	0	---	0	0	----
31	0	0	---	0	0	----
32	1	12	8.33	0	0	----
33	0	0	---	0	0	----
34	0	0	---	0	0	----
35	0	0	---	0	0	----
36	0	11	0.0	0	0	----
37	0	0	---	0	0	----
38	0	0	---	0	0	----
≥39	0	14	0.0	0	24	0.0
≤21	22	1731	1.27	12	271	4.43
≥22	9	471	1.91	2	181	1.10

For any category of sex and phrases required to enroll greater than 26, there is no more than one error and no more than 35 entry attempts. This is not enough data to be statistically significant. Those people who were re-enrolled or should have been re-enrolled required 21, 21, 26, 20, 22, and 21 for their initial enrollment. These people did not require any more phrases to enroll than those who did not require re-enrollment.

Since there is not enough data in each category of phrases required to do a comparative analysis, the data was combined into two categories: less than or equal to 21 phrases required and greater than 21 phrases required as shown at the bottom of Table VII and VIII. Males and females were combined in applying the F test to yield:

$$T = \frac{(1744 + 271)}{(475 + 183)} \cdot \frac{(13 + 4) + 1}{(35 + 12) + 1}$$

$$= \frac{2015}{658} \cdot \frac{18}{48} = 1.15$$

$$F(96,36) = 1.50$$

Similar results are obtained for Set 2 data. Therefore, it cannot be said that the error rates of the two groups are significantly different.

4.2.2.7 Type I Error Rate Versus Day Of Week. Type I error rate versus the different days of the week as shown in Table

IX showed no consistent pattern. One Monday during the test was a holiday and during that week, Tuesday was the first work day of the week. If the data from that Tuesday is included in the Monday column, the results are as shown in Table IX. These results include all the data used in Table II and show the daily variation (on the average) from the 2.32% total Type I error rate. The error rate does not appear larger on the day following a weekend or on any other particular day.

TABLE IX

TYPE I ERROR RATE VERSUS DAY OF WEEK

	Mon	Tues	Wed	Thurs	Fri
Normal Week Error Rate (%)	1.89	2.89	2.83	2.01	2.79
Holiday Week Error Rate (%)	2.43	2.41	2.83	2.01	2.79

4.2.3 Objective 3 - Independence of Type I Scores

To determine the independence of the Type I scores for repeated uses of the system by individual enrollees as well as when compared against other enrollees.

4.2.3.1 Independence Of Type I Errors Versus Individuals. Of the 213 people enrolled in the ASV system, 209 had one or more entry attempts. Most had no Type I errors at all. Table X shows the number of people who had the indicated number of errors.

TABLE X

TYPE I ERROR VERSUS INDIVIDUALS

Type I Errors (N)	Number of People with N Type I Errors	Total Errors
0	176	0
1	23	23
2	6	12
3	2	6
4	1	4
5	1	5

A minority of those using the system (15.8%) accounted for all of the Type I errors. Five percent of all the people using the system accounted for more than half of the Type I errors. Thus, it appears that the Type I errors are not uniformly distributed among all users of the system.

4.2.3.2 Distribution Of Type I Decision Function Scores. The speaker verification system asks a person to repeat a phrase. The ASV algorithm then computes a decision function score and compares it with a threshold. If the decision function score is less than the threshold, the person is permitted to enter, but if it is not, the person is asked to repeat a second phrase. The second phrase decision function score along with the first score and a second threshold are used to make a decision. A person who repeats four phrases gets a decision function score for his first, second, third, and fourth phrase. Occurrences of a decision function score

(i.e., phase 1, 2, 3 or 4) versus the value of the decision function score could be plotted for each phrase, but the plot for the earlier phrases would have many more entries than the later phrases since people who verify on the earlier phrases never say the later phrases.

To compare the distribution of occurrence versus decision function scores for the first, second, third, and fourth phrases, the number of occurrences of a given decision function score for a given phrase is divided by the number of times that phrase is used. This yields a probability density function (pdf). The resulting four pdf's are plotted in Figure 4.

The functions have virtually the same mean, same standard deviation, and same shape in the tails. Thus, the decision function scores appear to be taken from the same pdf whether it is the first, second, third, or fourth phrase spoken in any entry attempt. Since the decision function is used to determine verification, the proportion of people failing to pass on any given phrase would be independent of the phrase if each phrase used the same threshold and if previous decision function scores were not averaged with the present score. That is, people requiring more than one phrase to verify do not have a higher probability of being rejected when the same threshold is used and previous scores are not averaged with the present score. The algorithm used does vary the threshold and average the scores (cumulative decision strategy) because Type II errors must also be controlled. (See 4.2.7.1.)

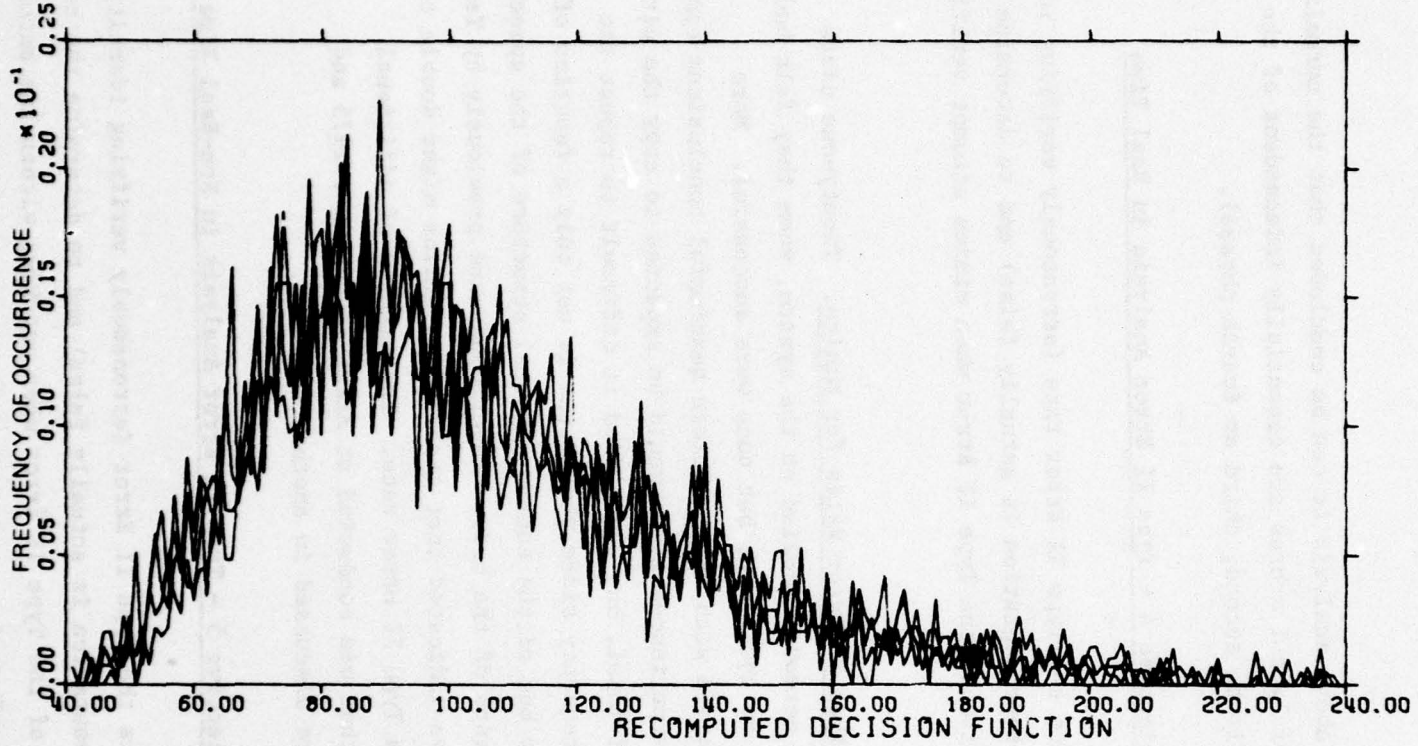


FIGURE 4 FREQUENCY OF OCCURRENCE VS. RECOMPUTED DECISION FUNCTION TYPE I

From the above analysis it can be concluded that the normalized distribution of Type I scores are essentially independent of the phrase said (first, second, third or fourth phrase).

4.2.4 Objective 4 - Type II Error Analysis in Real Time

To estimate the Type II Error rate (erroneously verifying identity when the representation is actually false) and to determine the confidence limits of the Type II Error when mimics attempt verification.

4.2.4.1 Type II Error Rates For Mimics. Twenty-one users tried to mimic persons enrolled on the system, whom they felt had voices similar to their own, but none were successful. More testing is required with mimics before meaningful conclusions can be drawn. An experienced mimic would be expected to copy the pitch of a given individual, but would find it difficult to repeat the same formant frequency since the latter is not only a function of the vocal cords but of the entire physical structure of the speech-producing elements of the body. Tests conducted previously by Texas Instruments have indicated that experienced mimics might double the observed random Type II error rate. The results of additional mimic testing that was conducted at MITRE in December 1975 and October 1976 are discussed in another document.

4.2.5 Objective 5 - Type II Error Analysis in Non-Real Time

To estimate the Type II Error (erroneously verifying identity when the representation is actually false) and to determine the confidence limits of the Type II Error when matching within the enrolled population is performed.

4.2.5.1 Type II Errors. The Type II error testing conducted is of a random nature. That is, it is of the form of entering someone else's identity number, either accidentally or intentionally, and responding to the prompted phrase in the usual manner without changing the voice, pitch, accent or speed. This data was generated by collecting enough raw data (eight phrases) during routine real time data collection, and then playing this data against all other reference files in non-real time. Use of a common vocabulary by all users makes this type of testing meaningful. The Type II error rates are shown in Table XI. Post Enrollment means that the victim i.e., the reference file, is in Post Enrollment. Normal means the victim is in the Normal verification state.

TABLE XI

TYPE II ERROR RATES

Post Enrollment	Errors	Attempts	Error Rate (%)
Male vs. Male	72	15,058	0.48
Mixed	1	5,569	0.02
Female vs. Female	4	518	0.77
Total	77	21,145	0.36
Normal			
Male vs. Male	418	42,128	0.99
Mixed	5	18,412	0.03
Female vs. Female	37	1,909	1.94
Total	460	62,449	0.74

It is unlikely that any intruder would attempt to enter a secure area by claiming the identity of someone of the opposite sex. Thus, the monosexual male versus male and female versus female error rates are considered to be more important.

To determine if the ASV Type II error rates for male intruder against male reference file and for female intruder against female reference file are significantly different, the F ratio test is used. In PE:

$$T = \frac{15058}{518} \cdot \frac{5}{73} = 1.99$$

$$F(146,10) = 2.08$$

and in Normal operation,

$$T = \frac{42128}{1909} \cdot \frac{38}{419} = 2.00$$

$$F(838,76) = 1.26$$

Therefore, in PE, it cannot be said that the female against female Type II errors are significantly different at the 90% confidence level than the male against male Type II errors, but in Normal operation they are significantly different. It is not clear why this difference in Normal operation exists. The ESE was shown to be higher for females than for males (see 4.2.5.3 for the effect), but this was true both in PE and Normal.

To compare the PE error rate with the Normal error rate, the male vs. male and female vs. female errors and attempts have been combined to yield:

$$T = \frac{15576}{44037} \cdot \frac{456}{77} = 2.09$$

and

$$F(154,912) = 1.16$$

so that the Type II error rate in PE is significantly different than it is in Normal operation. This is as expected since all four phrases must register in PE before a pass decision is made. In Normal operation, a decision can be made after any registered phrase.

4.2.5.2 Type II Errors Versus Entry Trials. 191 people had 4 or more trials and 107 people had 10 or more trials. The number of Type II errors, the number of entry attempts, and the error rate are tabulated versus trial for each of these groups in Tables XII and XIII.

As noted earlier, the first four successful verifications use a different strategy than do the subsequent verifications. In PE (Table XII, and trials 1 - 4 in Table XIII) four registered phrases are always required for a verification to occur. Thereafter, a decision can be made on any phrase from the first to as many as ten. As seen in 4.2.5.1, the Type II error rate in PE is lower than it is in Normal operations. This is also true for the 107 files that contributed to Table XIII. This is what was expected from the different strategies used in PE and Normal verification.

TABLE XII

TYPE II ERRORS FOR REFERENCE FILES WITH AT LEAST 4 TRIALS

Trial	1	2	3	4
Errors	8	17	27	43
Attempts	4574	9002	9194	11664
Error rate (%)	0.175	0.189	0.294	0.369

4.2.5.3 Type II Error Rate Versus Expected Scanning Error ($\hat{E}SE$). Type II error rate versus ESE is shown for the first 150 enrollees only (because of data processing limitations) in Figure 5. The error rate ordinate is composed of male vs. male, female vs. female, and mixed results.

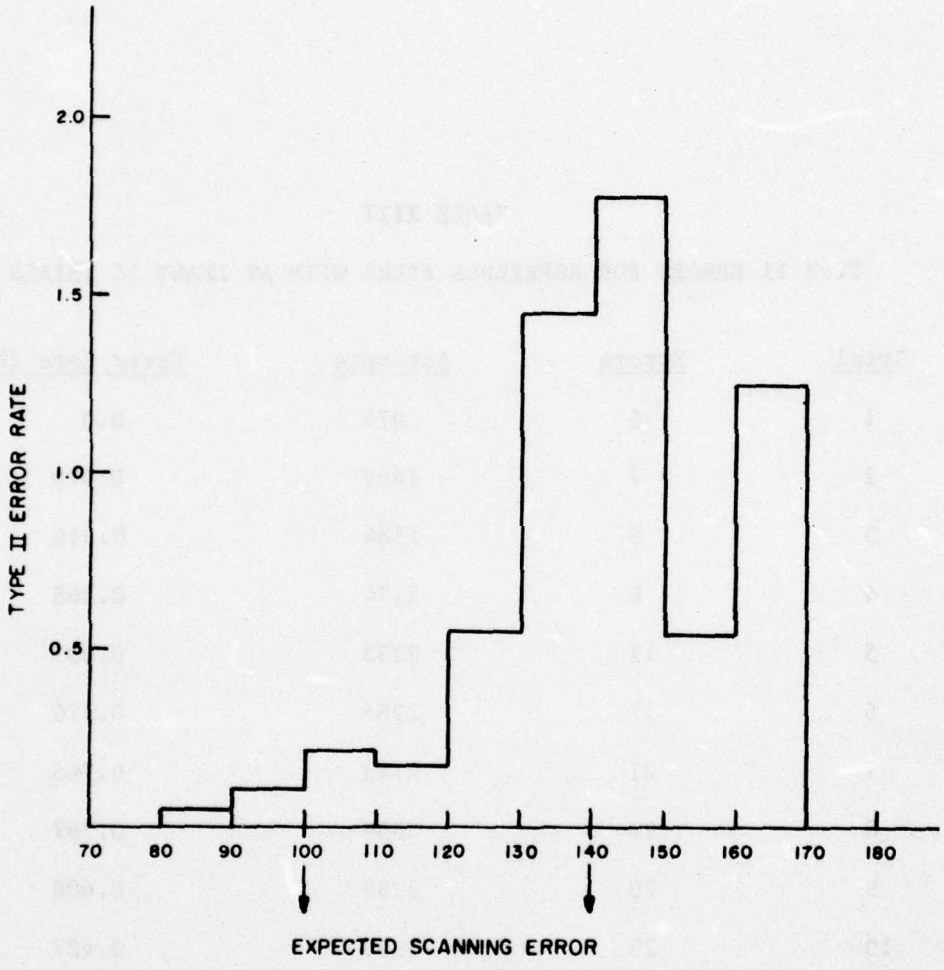
The data shows that Type II error rates increase dramatically with increasing expected scanning error of the reference file. This means that the people with high ESEs have higher Type II error rates and that Type II error rates are not uniformly distributed among individuals. This is in opposition to the Type I error rate performance, 4.2.2.3, where the error rate did not appear to depend on the ESE.

4.2.5.4 Type II Error Rate Versus Phrases Required To Enroll. The Type II error rate versus phrases required during enrollment for male-male and female-female Type II attempts are given in Tables XIV and XV, respectively.

TABLE XIII

TYPE II ERRORS FOR REFERENCE FILES WITH AT LEAST 10 TRIALS

Trial	Errors	Attempts	Error Rate (%)
1	0	974	0.0
2	7	1489	0.470
3	5	1584	0.316
4	8	2174	0.368
5	11	2233	0.493
6	17	2985	0.570
7	21	2743	0.766
8	28	3559	0.787
9	20	3285	0.609
10	23	5390	0.427



IA-50,455

Figure 5 TYPE II ERROR RATE VS EXPECTED SCANNING ERROR
(FOR FIRST 150 PEOPLE ENROLLED ON ASV SYSTEM)

TABLE XIV

TYPE II ERROR RATE VERSUS PHRASES TO ENROLL (MALE-MALE)

Phrase Required During Enrollment	Errors	Attempts	Error Rate (%)
20	201	27879	0.72
21	93	13470	0.69
22	63	2509	2.51
23	47	3147	1.49
24	4	1065	0.38
25	40	4191	0.95
26	10	1735	0.58
27	16	368	4.35
28	7	372	1.88
29	7	741	9.46
30	0	0	——
31	0	375	0.0
32	0	352	0.0
33	0	0	——
34	0	0	——
35	0	0	——
36	0	374	0.0
37	0	0	——
38	0	0	——
$\geq$ 39	2	608	0.33
$\leq$ 21	294	41349	0.71
$\leq$ 22	196	15837	1.24

TABLE XV

TYPE II ERROR RATE VERSUS PHRASES TO ENROLL (FEMALE-FEMALE)

Phase Required During Enrollment	Errors	Attempts	Error Rate (%)
20	12	763	1.57
21	0	505	0.0
22	4	340	1.12
23	0	42	0.0
24	0	0	—
25	8	250	3.20
26	12	269	4.46
27	0	0	—
28	2	138	1.45
29	0	0	—
30	0	0	—
31	0	0	—
32	0	0	—
33	0	0	—
34	0	0	—
35	0	0	—
36	0	0	—
37	0	0	—
38	0	0	—
$\geq$ 39	3	120	2.50
$\leq$ 21	12	1268	0.95
$\geq$ 22	29	1159	2.50

The Type II error rates do not show any trend with varying phrases required during enrollment. However, applying the F-test to those reference files requiring 22 or more phrases and those requiring 20 or 21 phrases to enroll yields:

$$T = \frac{41349}{15837} \cdot \frac{197}{295} = 1.74$$

$$F(590,394) = 1.00$$

Therefore, those reference files requiring 22 or more phrases have a significantly different error rate than those requiring 20 or 21 phrases to enroll.

4.2.5.5 Speaker Average Versus Number of Phrases Required to Enroll. Table XVI presents the reference file speaker average at the end of enrollment versus the number of phrases required to enroll. The category less than or equal to 21 phrases and the remainder as 22 or more are also presented.

As noted in 4.2.2.3 and 4.2.2.6, the Type I error rate did not increase with number of phrases required to enroll, nor with increasing speaker average (expected scanning error). This indicates that the normalization process used prior to making a pass/ fail decision is effective. Thus, even though those people requiring more than 21 phrases to enroll have a higher speaker average, their Type I error rate is not significantly different.

Conversely, 4.2.5.3 and 4.2.5.4 show that the Type II error rate does increase with phrases required to enroll and speaker average (ESE). The vulnerable points of the ASV system to intruders,

TABLE XVI

SPEAKER AVERAGE VERSUS PHRASES TO ENROLL

No. Phrase To Enroll	Males		Females	
	No. Cases	Speaker Average	No. Cases	Speaker Average
20	83	112	13	146
21	40	114	8	135
22	8	125	5	156
23	9	125	1	109
24	3	126	0	—
25	13	121	4	157
26	5	129	4	162
27	1	140	0	—
28	1	152	2	135
29	2	129	0	—
30	0	—	0	—
31	0	—	0	—
32	1	136	0	—
33	0	—	0	—
34	0	—	0	—
35	0	—	0	—
36	1	104	0	—
37	0	—	0	—
38	0	—	0	—
$\geq$ 39	3	126	2	139
$\leq$ 21	123	113	21	142
$\geq$ 22	47	125	18	151

therefore, are those reference files with higher speaker averages. From Table XVI, it appears that phrases to enroll is an indicator of high speaker average, but that the only sure way of identifying those files is by looking at the speaker average itself. The number of phrases to enroll is only an indicator of the degree of difficulty that the enrollee had.

4.2.6 Objective 6 - Independence of Type II Scores

To determine the independence of the Type II scores for repeated uses of the system by individual enrollees as well as when compared against other enrollees.

4.2.6.1 Distribution Of Type II Decision Function Scores. A distribution of frequency of occurrence versus decision function score can be plotted for intruders who registered against a reference file. Registering against a reference file implies meeting the prescreening requirements which is a gross correlation between reference file and speaker data prior to performing a test against the decision function threshold for a pass/fail decision. To compare the distribution of occurrence for the first phrase said to the distribution for succeeding phrases, each distribution is divided by the total number of times that the phrase registered. The normalized distributions of frequency of occurrence versus decision function score are plotted one on top of the other for each phrase in Figure 6. (See 4.2.7.1.)

The distributions are fairly similar -- the mean and standard deviation of each distribution moves slightly higher with increasing phrase number. This increase of the mean is probably due to the

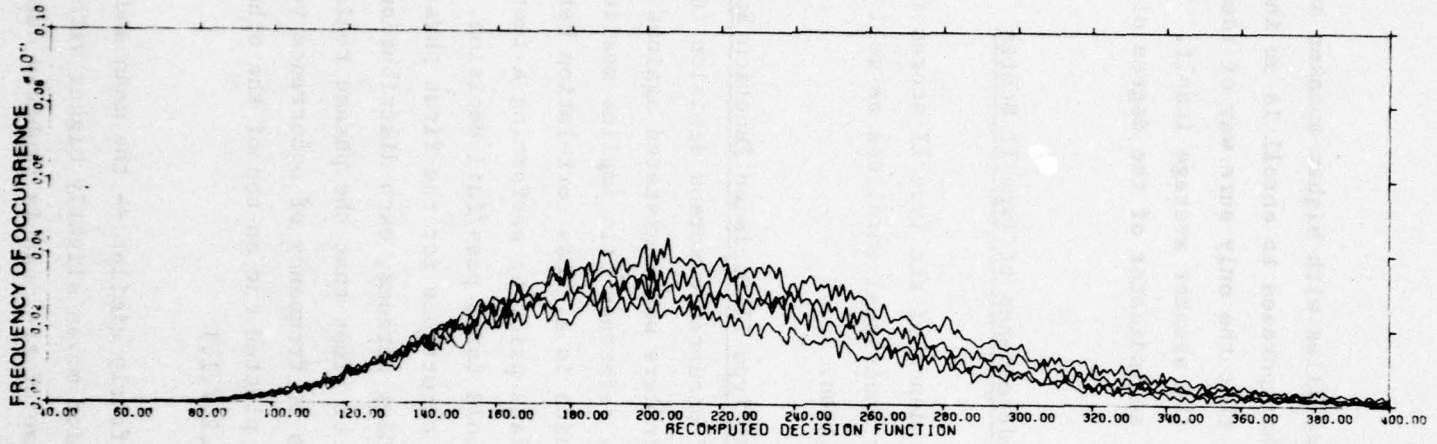


FIGURE 6 FREQUENCY OF OCCURRENCE VS. RECOMPUTED DECISION FUNCTION TYPE II

selection process which drops the intruders with low decision function scores because they get a Type II error and do not try on the succeeding phrases. This leaves intruders who have slightly higher decision function scores. The occurrences shown here are not the total number of intruder attempts, but only those attempts in which the intruders speech pattern registered against the reference pattern.

4.2.7 Objective 7 - Sensitivity Analysis of Type I and Type II Errors to Thresholds

To determine the sensitivity of Type I and Type II Error variations to changing thesholds.

4.2.7.1 Sensitivity Analysis. The fraction of Type I Recomputed Decision Function Scores (RCDFS) greater than and the Type II RCDFS less than a particular value of the score (abscissa) are shown in Figures 7 - 10 for phrases 1-4, respectively. Figure 7 was obtained by integrating the curves in Figures 4 and 6 and presenting the results for phrase one from each in Figure 7. Figures 8, 9, and 10 for phrases 2, 3, and 4 were generated in an identical manner. The first curve in Figure 7 shows the fraction of all the Type I decision function values (decision functions for brevity) which were less than a particular decision function value. Type I decision functions were obtained for each verification attempt after a person keyed in his own identification (ID) number. This curve includes everyone, male and female, PE and Normal verification. Thus, of all the people who repeated the first prompted phrase, about 48% had a score higher than 100 and about 28% had a score higher than 120. The second curve in Figure 7 shows the fraction of

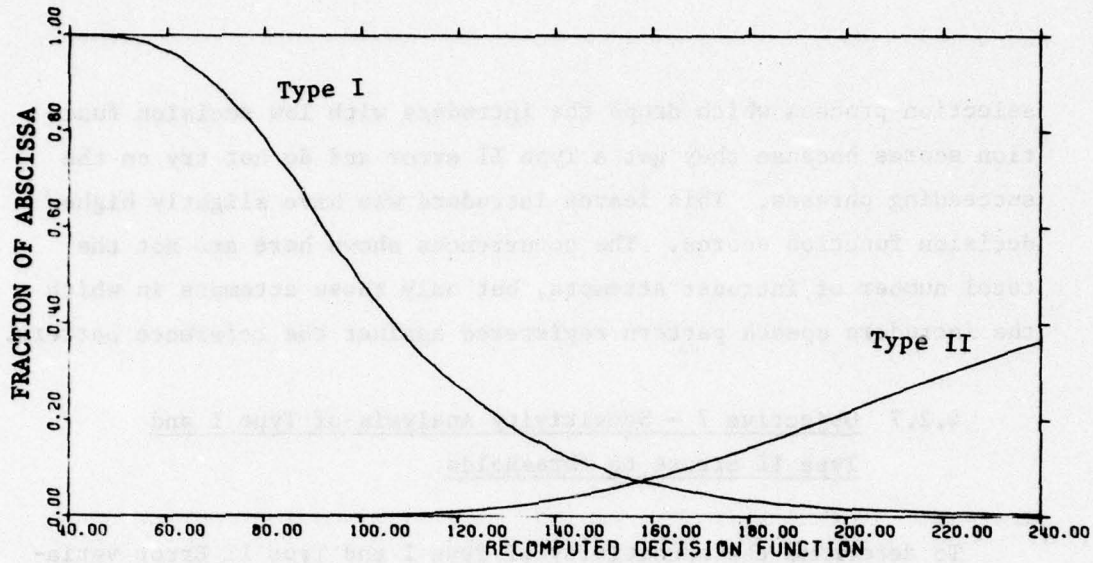


FIGURE 7 FRACTION OF TYPE I "RECOMPUTED DECISION FUNCTION SCORE" (RCDFS) GT AND TYPE II RCDFS LT ABSCISSA PHRASE 1

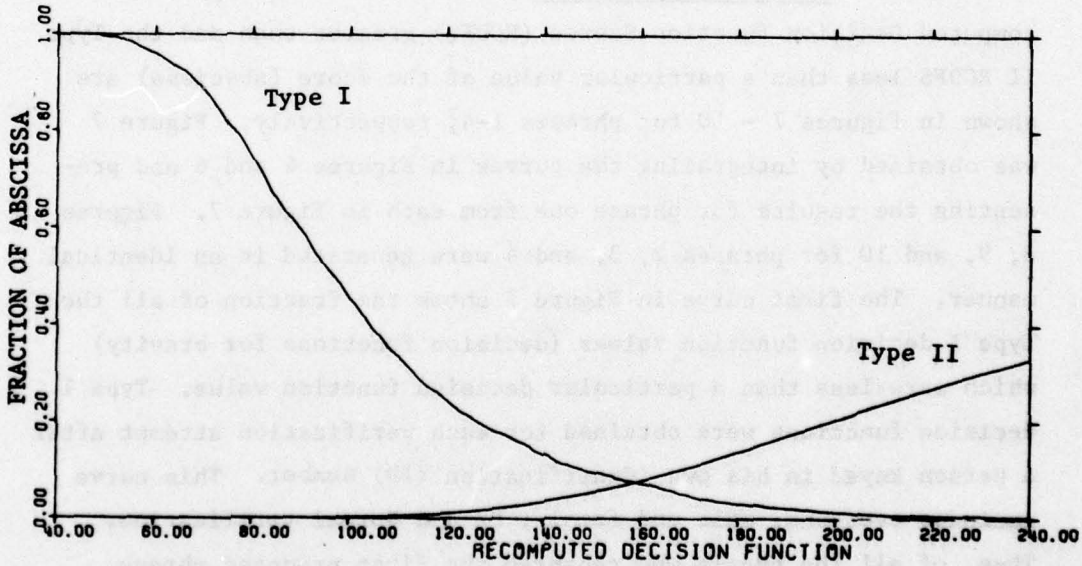


FIGURE 8 FRACTION OF TYPE I RCDFS GT AND TYPE II RCDFS LT ABSCISSA PHRASE 2

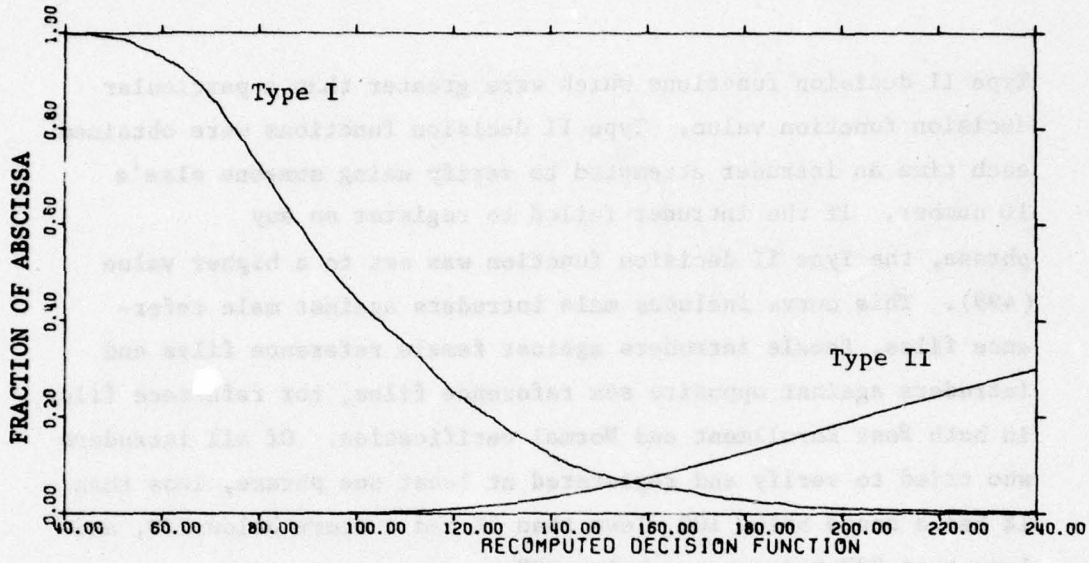


FIGURE 9 FRACTION OF TYPE I RCDFS GT AND TYPE II RCDFS LT ABSCISSA PHRASE 3

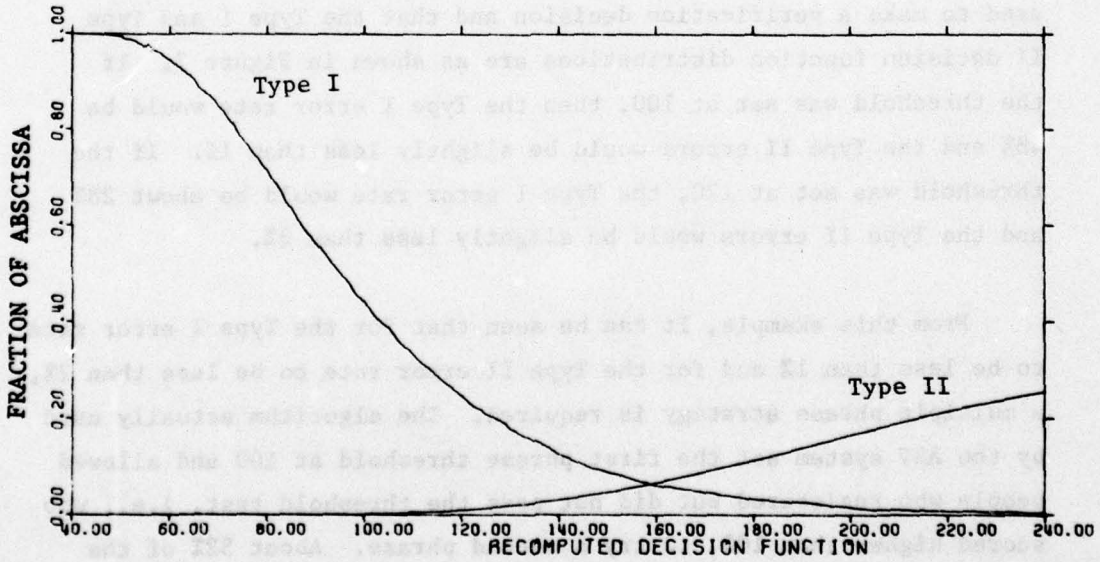


FIGURE 10 FRACTION OF TYPE I RCDFS GT AND TYPE II RCDFS LT ABSCISSA PHRASE 4

Type II decision functions which were greater than a particular decision function value. Type II decision functions were obtained each time an intruder attempted to verify using someone else's ID number. If the intruder failed to register on any phrase, the Type II decision function was set to a higher value (499). This curve includes male intruders against male reference files, female intruders against female reference files and intruders against opposite sex reference files, for reference files in both Post Enrollment and Normal verification. Of all intruders who tried to verify and registered at least one phrase, less than 1% had a score below 100, less than 3% had a score below 120, and less than 22% had a score below 200.

Suppose for a moment, that only a single four-word phrase is used to make a verification decision and that the Type I and Type II decision function distributions are as shown in Figure 7. If the threshold was set at 100, then the Type I error rate would be 48% and the Type II errors would be slightly less than 1%. If the threshold was set at 120, the Type I error rate would be about 28% and the Type II errors would be slightly less than 3%.

From this example, it can be seen that for the Type I error rate to be less than 1% and for the Type II error rate to be less than 2%, a multiple phrase strategy is required. The algorithm actually used by the ASV system set the first phrase threshold at 100 and allowed people who registered but did not pass the threshold test, i.e., who scored higher than 100, to try a second phrase. About 52% of the people using their own ID passed on the first phrase; the other 48% tried a second phrase. Less than 1% of all intruders had scored less

than 100. Thus, Type II errors are less than 1% after the first phrase and over 99% of the intruders must try at least one additional phrase. Curves of the distribution of Type I decision function values and of Type II decision function values for the second phrase are shown in Figure 8. The threshold for phrase 2 was set at 120 for the test. Figure 8 shows that this setting resulted in a 2% Type II error rate. About 74% of the Type I decision function values were less than 120. This means that of those people required to speak a second phrase, about 74% passed on the second phrase. The other 26% were required to speak a third phrase. Curves for the third and fourth phrases are shown in Figures 9 and 10 where the threshold were set at 135 and 145, respectively.

Table XVII shows the percentage of people using their own ID who verified and the percentage of registered phrases for intruders versus phrase number. For this hypothesized four phrase strategy, the Type I error rate is only 0.2%, but the Type II error rate in Table XVII is between 4% and 10% (for this hypothesized case, the average number of phrases for authorized users to verify is 1.62). If one assumes that the 4% who scored less than 145 are the same group who scored less than 135, and the 2% who scored less than 120, and the 1% who scored less than 100, then the best four phrase error rate is 4%. On the contrary, if the 2% who scored less than 120 are a different group from those who scored less than 100, etc., then a worst Type II error rate of 10% is arrived at.

Since the number of cases with 1, 2, 3, or 4 registered phrases varies, the number of Type II errors also varies for each phrase. In addition, there were 34,959 cases where none of the eight phrases registered so that there were actually 83,594

TABLE XVII

HYPOTHESIZED PERFORMANCE BASED ON FIGURES 7 - 10

Phrase Number	Threshold	TYPE I			TYPE II		
		% Decision Function Threshold	% Total Passes After Phrase	Resulting Type I Error Rate (%)	% Decision Function Threshold	% Total Passes After Phrase	
						Best	Worst
1	100	52	52.0	48.0	1	1	1
2	120	74	87.5	12.5	2	2	3
3	135	84	98.0	2.0	3	3	6
4	145	90	99.8	0.2	4	4	10 –

Type II attempts (see Table XVIII). This results in the best and worst case estimates of the Type II error rate to be 1.56% and 3.34%. This treatment of misregistered phrases more nearly follows the speaker verification strategy.

TABLE XVIII

HYPOTHESIZED TYPE II ERROR RATES INCLUDING MISREGISTERED PHRASES

Number of Registered Phrases	Number of Cases	Type II Errors Based on Table XVII Best	Worst
1	48,635	486	486
2	35,139	351	703
3	27,001	270	810
4	19,823	198	793
Total Errors*	83,594	1,305	2,792

*Includes 34,959 cases of misregistered phrases.

The correct way to do a sensitivity analysis is to: (a) decide on the new desired performance, (b) use all conditions that the strategy does, (c) adjust the thresholds or other equivalent parameters in the program, (d) reprocess recorded data to get a confirmation of the expected performance, and (e) collect new data, if necessary, to establish the performance.

Adjusting any of the four thresholds to the right will increase the Type II error rate, but will decrease the Type I error rate. Adjusting the threshold to the left will have the opposite effect, that is, decreased Type II and increased Type I error rates. This

strategy operates differently from the one used in the ASV system but it does give an example of some of the initial analysis required to start a sensitivity analysis.

The three primary differences between the above strategy and the ASV strategy are in the computation of the decision function (DF), the use of conditional events and the treatment of misregistered phrases. In ASV, the DF is computed as:

$$DF_N = \frac{\sum_{i=1}^{N} SE_i}{100N < \sum_{i=1}^{N} ESE_i < 140N}, \quad N = 1, 2, 3, 4$$

where SE_i is the scanning error and ESE_i the expected scanning error for phrase i. The functions computed in Figures 7 through 10 are on a per phrase basis without memory for the first four registered phrases, and the limiting in the denominator was not used. The conditional probabilities are important especially for the Type II data. That is, given that phrase 1 misregistered, what is the probability that phrase 2 will also misregister. Also, misregistered phrases are by themselves a discrimanent. In PE, three misregistered phrases and in Normal, two misregistered phrases caused the four phrase strategy to be terminated and a decision to be made.

4.2.8 Objective 8 - Verification Time Analysis

To determine the average time required for verification and the variance about this time.

4.2.8.1 Service Time (Verification Time). Long delays in verifying identity and getting into a secure area will be unacceptable to the entrant and perhaps to management. Service time is made up of keyboard time, verification time, first door opening-closing, second door opening-closing time, and dead time. These last three times will not be discussed here. The average time to stroke four digits at the keyboard, hear the spoken digits and hit the SEND key was 3.24 seconds. The standard deviation about this time was 1.75 seconds. Table XIX shows a breakdown of the number of decisions, including those who did not verify, by phrase number. During PE, an average of 4.77 phrases were required to verify while in Normal verification an average of only 1.70 phrases were required to verify. The average response time per phrase was 1.948 seconds. The standard deviation about this time was 0.41 second. The time required to prompt a phrase was 1.9 seconds. The sum of this time and the average response time multiplied by the number of phrases yields the average verification time. This time was 18.35 seconds during PE but only 6.54 seconds during Normal verification. Thus, after an individual has four successful verification attempts and is in Normal verification, the average time for keyboard and verification is 9.8 seconds. In an operational system only a small number of users would be new to the installation and be in PE at any one time. A discussion of throughput is presented in Volume V of this report.

TABLE XIX

NUMBER OF DECISIONS VERSUS PHRASE NUMBER

Phrase Number	1	2	3	4	5	6	7	8	9	10	11	12
Post Enrollment	-	-	-	649	68	20	10	53	21	11	12	6
Normal	1054	500	182	49	4	3	5	29	-	-	-	-

5.0 PHASE II TEST

5.1 DESCRIPTION

The following is a description of the Phase II test setup and the changes made to the ASV system following the completion of Phase I. The ASV algorithm used in the Phase II test was the same as the Phase I ASV algorithm except for four modifications. These modifications were made in an attempt to reduce the high Type I error rate and increase the low Type II error rate that occurred in the PE portion of Phase I.

5.1.1 Test Setup

The Phase II test was setup similar to that for Phase I discussed in 4.1. The data collection period went from 16 August through 8 October 1976. The test population consisted of 200 people enrolled but only 199 people (164 male and 35 female) used the system following enrollment. Of the total population enrolled, 99% returned for at least one verification. Most of the users had also participated in Phase I. Each user was requested to verify each day of the test, but no more than twice in any one day.

5.1.2 Noise Normalization Change

The noise normalization equation was changed from the Phase I form of:

$$x_{ij} = \frac{x_{ij}}{u_j}$$

to

$$x_{ij} = \frac{x_{ij} - u_j}{s_j}$$

where

j = index on time

i = 1, ..., 14 index on the 14 filters

u_j = mean power across the bandwidth in time sample j

s_j = standard deviation of the power across the bandwidth in time sample j

x_{ij} = power at the output of the i^{th} filter at time j

5.1.3 End of Enrollment Speaker Average Estimate Change

The estimate of speaker average after enrollment was changed from the Phase I form of:

$$ESE_i = b_i \cdot ESE_{ei}$$

to

$$ESE_i = \text{Max}\ (140, b_2 \cdot ESE_{ei})$$

where

ESE_{ei} = the estimated speaker average at the end of enrollment for each of the 16 words

ESE_i = the estimated speaker average including an estimate of day to day variability

$i = 1, 2, \ldots, 16$

$b_1 = 1.17$

$b_2 = 1.25$

5.1.4 Decision Function Calculation Change

The decision function calculation is

$$DF_N = \sum_{i=1}^{N} SE_i \Big/ \left(100N < \sum_{i=1}^{N} ESE_i < max_i \cdot N \right)$$

For Phase I the value of max_i was always 140.

For Phase II the values of max_i was changed to:

max_i = 160 for Post Enrollment

max_i = 130 for phrase 1 in Normal (N=1)

= 135 for phrase 2 in Normal (N=2)

= 140 for phrase 3 in Normal (N=3)

= 145 for phrase 4 in Normal (N=4)

N = 1, 2, 3, or 4

(and repeats in auto abort.)

5.1.5 Phase Recycling in Normal Mode

In Phase II, recycling (defined in 5.2.4.1) was included in Normal whereas no recycling in Normal verification was used in Phase I. Post Enrollment had recycling in both phases.

5.2 RESULTS

5.2.1 Objective 1 - Type I Error Analysis in Real Time

To estimate the Type I Error rate (failure to verify proper identity when the representation is actually true) and to determine the confidence limits of the Type I Error estimate.

5.2.1.1 Type I Errors. Table XX shows the Type I errors, trials, and error rates for all users which occurred in both PE and Normal processing. There were a total of 17 Type I errors of which 3 had assignable causes. The assignable causes were eating candy (1), harassment (1), and an intentional fail for demonstration purposes (1). The assignable errors have been removed from any consideration in all of the following analysis except Table XX and XXI Set 1, where only the intentional fail is not considered. Five of the Type I errors were made by people who had colds.

To determine if the PE error rate is significantly different from the Normal verification error rate, the F ratio test is used:

$$T = \frac{4252}{768} \cdot \frac{3}{15} = 1.11$$

TABLE XX

TYPE I ERROR RATES - ALL USERS

	With Assignable Causes		
	Errors	Attempts	Error Rate (%)
Post Enrollment	2	768	0.26
Upper Bound*			0.69
Normal	14	4252	0.33
Upper Bound*			0.47
Total	16	5020	0.32
Upper Bound*			0.45

	Without Assignable Causes		
	Errors	Attempts	Error Rate (%)
Post Enrollment	2	768	0.26
Upper Bound*			0.69
Normal	12	4245	0.28
Upper Bound*			0.42
Total	14	5013	0.28
Upper Bound*			0.38

*90% confident that the true error rate is less than the upper bound. This is the Chi-square test. See Appendix B.

$$F_{p=.9}(30,6) = 2.80$$

Therefore, it cannot be said at the 90% confidence level that the PE error rate is significantly different than the Normal verification error rate. This result is most likely due to the changes made to the algorithm (see 5.1). This was a goal of the changes made to the algorithm, and it was apparently successful.

Both the best estimate of the error rates and the 90% confidence level on the upper bound of the error rates are well below the maximum acceptable Type I error rate of 1%.

5.2.2 Objective 2 - Type I Error Analysis in Non-Real Time

To determine how various parameters affect the Type I Errors and system performance.

5.2.2.1 Type I Errors Versus Sex. Table XXI tabulates the Type I error rates versus sex. Set 1 contains all assignable errors while Set 2 has the two assignable errors removed. (See 5.2.1.1.) From the table, the error rates for both males and females are all less than the maximum acceptable rate of 1%. For this sample, the data shows that the males have a lower Type I error rate than females during both PE and Normal verification.

To determine if the male and female error rates are significantly different the F ratio test is used.

TABLE XXI

TYPE I ERROR RATES VS. SEX

Set 1

	Errors	Attempts	Error Rate (%)
Males P.E.	1	623	0.16
Males Normal	7	3456	0.20
Females P.E.	1	145	0.69
Females Normal	7	796	0.88

Set 2

	Errors	Attempts	Error Rate (%)
Males P.E.	1	623	0.16
Males Normal	7	3456	0.20
Females P.E.	1	145	0.69
Females Normal	5	794	0.63

For PE, the F test yields:

$$T = \frac{623}{145} \cdot \frac{2}{2} = 4.30$$

$$F(4,4) = 4.11$$

For Normal verification, the F test yields:

$$T = \frac{3456}{794} \cdot \frac{6}{8} = 3.26$$

$$F(16,12) = 2.09$$

Thus, the male and female Normal verification error rates are significantly different in both PE and Normal verification.

However, it cannot be said that the error rates in PE for males are significantly different from those in Normal verification as shown below.

$$T = \frac{3456}{623} \cdot \frac{2}{8} = 1.39$$

$$F(16,4) = 1.98$$

This result is also true for females since:

$$T = \frac{794}{145} \cdot \frac{2}{6} = 1.82$$

$$F(12,4) = 3.90$$

Comparing total male errors to total female errors yields:

$$T = \frac{4079}{939} \cdot \frac{7}{9} = 3.38$$

$$F(18,14) = 1.98$$

Therefore, the combined Normal and PE error rate for males is significantly different than those for females. The females had higher error rates than males.

5.2.2.2 Type I Errors Versus Time Of Day. Type I errors versus time of day is shown in Table XXII. The error rates are much smaller than the desired maximum error rate of 1%.

The F test for morning and afternoon data yields:

$$T = \frac{2325}{2693} \cdot \frac{11}{5} = 1.89$$

$$F(10,22) = 1.90$$

Therefore, at the 90% confidence level it cannot be maintained that the morning error rate is statistically different from the afternoon error rate.

TABLE XXII

TYPE I ERROR RATES VERSUS TIME OF DAY

	Errors	Attempts	Error Rate (%)	Upper Bound (%)
Morning	10	2693	0.37	0.57
Afternoon	4	2325	0.17	0.34

5.2.2.3 Type I Error Rates Versus Expected Scanning Error.* The expected scanning error (ESE) is a measure of the consistency between repetitions of the same words. The lower the ESE, the better the match between reference patterns and new speech material. The ESE is used to normalize the current scanning error and to determine a decision function which is compared against a threshold.

*Expected scanning error and speaker average refer to the same quantity and are used interchangeably in this report.

Figure 11 shows the Type I error rate versus ESE for the combined data of males and females in both Normal and PE. This figure was derived using the value of ESE that each individual had at the end of the test period. For each 10 units of ESE the Type I error rate was calculated by summing the individual errors and dividing by the sum of the individuals attempts. For values of ESE less than 160 where greater than 90% of the population had values of ESE, the Type I error trend remains approximately flat. This indicates that Type I errors do not depend on the ESE very strongly, as expected.

5.2.2.4 Type I Errors Versus Station. The ASV system in Phase II had two user terminals called Station 1 and Station 2 as in Phase I. All enrollments were at Station 1. Verifications could take place at either station, but because of the enrollment experience at Station 1 and the closer proximity of Station 1 to the entrance to the laboratory most took place at Station 1 (see Table XXIII).

TABLE XXIII

TYPE I ERROR RATES VERSUS STATION

	Errors	Attempts	Error Rate (%)	Upper Bound (%)
Station 1	6	3205	0.19	0.36
Station 2	8	1813	0.44	0.72

The error rates at both stations were well below the maximum acceptable rate. The F ratio test for Stations' 1 and 2 data yields:

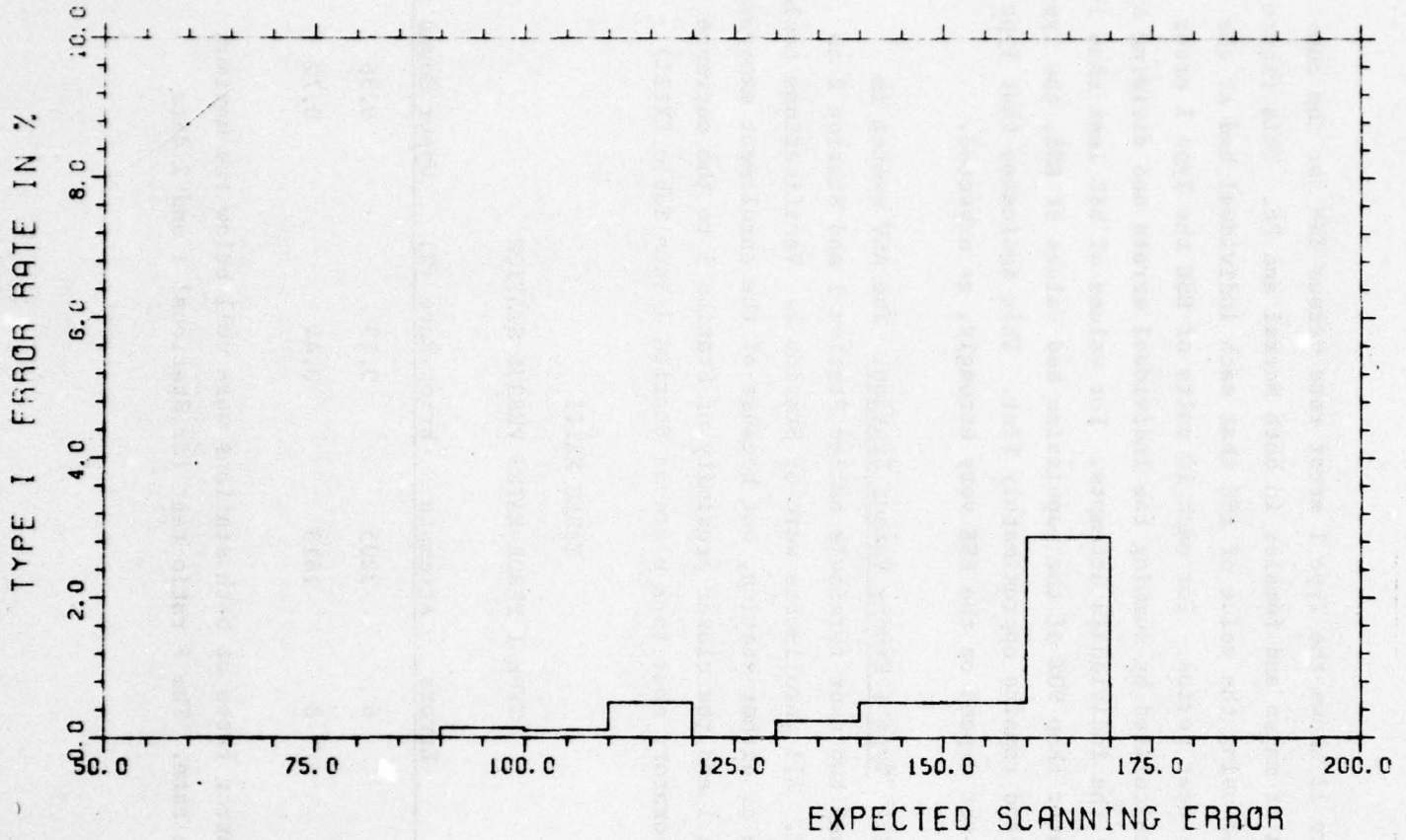


FIGURE 11 TYPE I ERROR RATE VERSUS EXPECTED SCANNING ERROR

$$T = \frac{3205}{1813} \cdot \frac{9}{7} = 2.27$$

$$F(14,18) = 1.89$$

Therefore, at the 90% confidence level, the error rate at Station 1 is significantly different than at Station 2. However, three of the total of eight errors occurring at Station 2 were caused by one person. This person had a cold the three times he attempted to verify and failed. One other Type I error due to a cold occurred at each station. The F test with the errors due to colds removed yields:

$$T = \frac{3204}{1809} \cdot \frac{5}{6} = 1.48$$

$$F(12,10) = 2.28$$

which changes the above result. Therefore, due to the large influence of one person with a cold, the results above are unreliable. A larger data sample is required to more accurately determine the correlation, if any, between stations.

5.2.2.5 Type I Errors Versus Entry Trials Since Enrollment. Because of the low number of Type I errors, no meaningful analysis of Type I errors versus entry trials since enrollment could be made.

5.2.2.6 Type I Errors Versus Phrases To Enroll. As in 5.2.2.5 an analysis was not made because of the low number of Type I errors.

5.2.2.7 Type I Error Rate Versus Day Of Week. Type I error rate versus the different days of the week as shown in Table XXIV showed no consistent pattern. One Monday during the test was a holiday and during that week, Tuesday was the first work day of the week. If the data from that Tuesday is included in the Monday column, the results are as shown in Table XXIV. These results include all the data used in Table XX (with assignable causes) and show the daily variation (on the average) from the 0.32% total Type I error rate.

TABLE XXIV

TYPE I ERROR RATE VERSUS DAY OF WEEK

	Mon	Tues	Wed	Thur	Fri
Normal Week Error Rate (%)	0.12	0.19	0.50	0.0	0.75
Holiday Week Error Rate (%)	0.10	0.23	0.50	0.0	0.75

5.2.2.8 Type I Errors Versus Personal Statistics. Tables XXV through XXVIII tabulate Type I errors as functions of the entrant's height, education level, age and primary education location. Since there are an insufficient number of users in each category, no reliable trend has been developed.

5.2.3 Objective 3 - Independence of Type I Scores

To determine the independence of the Type I scores for repeated uses of the system by individual enrollees as well as when compared against other enrollees.

TABLE XXV

TYPE I ERRORS VERSUS HEIGHT

Height (h) (inches)	Errors	Attempts	Error Rate (%)
h <61	1	150	0.67
61 ≤h <64	3	421	0.71
64 ≤h <67	6	937	0.64
67 ≤h <70	3	1567	0.19
70 ≤h <73	1	1465	0.07
73 ≤h	0	478	0

TABLE XXVI
TYPE I ERRORS VERSUS EDUCATION LEVEL

Years of School	Errors	Attempts	Error Rate (%)
≤ 8	0	0	0
<12	0	253	0
=12	4	717	0.56
<16	4	1011	0.40
≥16	6	3037	0.20

TABLE XXVII

TYPE I ERRORS VERSUS AGE

Age (Yrs.)	Errors	Attempts	Error Rate (%)
≤ 25	3	608	0.50
26-35	2	986	0.20
36-45	8	1496	0.53
46-55	1	1560	0.06
> 55	0	368	0

TABLE XXVIII

TYPE I ERRORS VERSUS PRIMARY EDUCATION LOCATION

Location	Errors	Attempts	Error Rate (%)
New England	8	2826	0.28
New York Area	5	1449	0.35
South	1	264	0.38
West	0	154	0
Total USA	14	4693	0.30
Foreign	0	325	0

5.2.3.1 Independence Of Type I Errors Versus Individuals. Of the 200 people enrolled in the ASV system 199 had one or more attempts. Table XXIX shows the number of people who had the indicated number of errors. Due to the low number of errors, no meaningful results can be established. For instance the one person with three errors due to colds dominates the results. This analysis would be more meaningful after every user had at least 200 trials. In Phase I the number of errors was much higher, thus making the results more meaningful.

TABLE XXIX

TYPE I ERROR VERSUS INDIVIDUALS

Type I Errors (N)	No. of People with N Type I Errors	Total Errors
0	191	0
1	9	9
2	1	2
3	1	3

Table XXX shows the number of people who had greater than 1, 3, and 5% Type I error rates. The number of errors these people had is tabulated. These numbers are indicated, also, as a percent of the total population (199 who had at least one trial) and as a percent of the total errors, respectively. As seen from the table, 5.5% of the people had 100% of the errors.

TABLE XXX

PERCENT OF TOTAL POPULATION AND TOTAL ERRORS AS FUNCTIONS OF THE TYPE I ERROR RATE

	Type I Error Rate		
	1%	3%	5%
No. of People	11	8	2
% of Total Population (199 who had 1 trial)	5.5%	4.0%	1.0%
No. of Errors	14	12	4
% of Total Errors (14)	100%	85.7%	28.6%

5.2.3.2 Distribution Of Type I Decision Function Scores. To compare the distribution of occurrence versus decision function score for the first, second, third, and fourth phrases, the number of occurrences of a given decision function score for a given phrase is divided by the number of times that phrase is used. As opposed to how the distribution of decision function scores were used in Phase I (4.2.3.2), these pdf's are exactly as used in the ASV algorithm. The resulting four pdf's are plotted in Figure 12. These plots are for all users in PE and Normal. The means of the distributions are 93.2, 103.5, 100.3, and 91.3 for phrases 1 through 4, respectively. The standard deviations of the distributions are 29.8, 25.7, 27.9, and 25.3 for phrases 1 through 4, respectively. This data is used directly in the sensitivity analysis (5.2.7.1).

Generally, one would expect the mean to rise in each succeeding phrase. There are two reasons that it does not: as noted in 5.1.4,

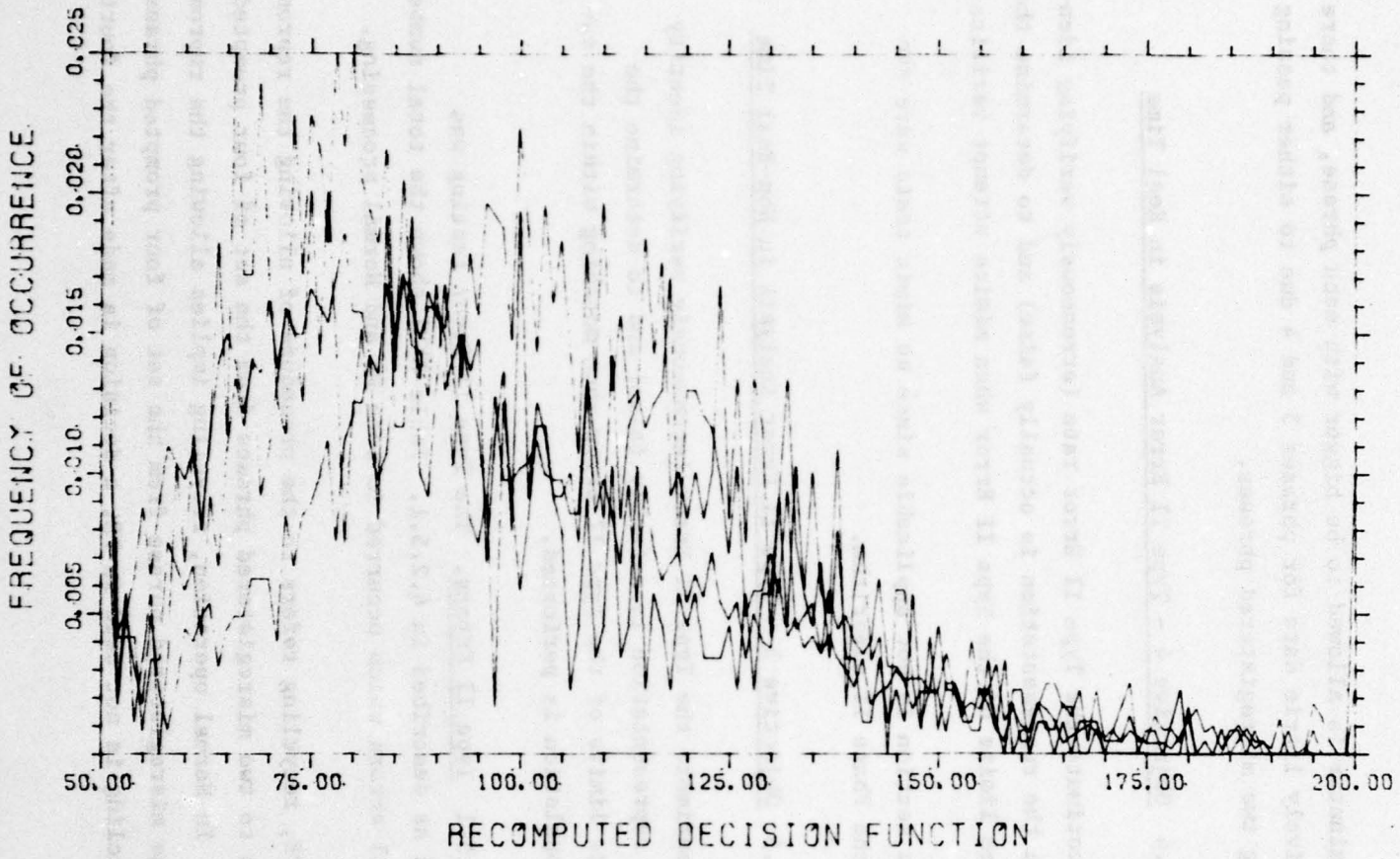


FIGURE 12 FREQUENCY OF OCCURRENCE VS. RECOMPUTED DECISION FUNCTION TYPE I

the denominator is allowed to be bigger with each phrase, and there is relatively little data for phrases 3 and 4 due to either passing or having two misregistered phrases.

5.2.4 Objective 4 - Type II Error Analysis in Real Time

To estimate the Type II Error rate (erroneously verifying identity when the representation is actually false) and to determine the confidence limits of the Type II Error when mimics attempt verification.

This section is not applicable since no mimic tests were run against the Phase II algorithm.

5.2.5 Objective 5 - Type II Error Analysis in Non-Real Time

To estimate the Type II Error (erroneously verifying identity when the representation is actually false) and to determine the confidence limits of the Type II Error when matching within the enrolled population is performed.

5.2.5.1 Type II Errors. The Type II Error testing was conducted as described in 4.2.5.1. Table XXXI shows the total number of Type II errors which occurred in both PE and Normal processing.

In PE, recycling refers to the procedure of allowing the reprompting of up to two misregistered phrases from the set of four prompted phrases. In Normal operation, recycling implies allowing the reprompting of one misregistered phrase from the set of four prompted phrases. When recycling is not used in PE, a decision is made after the fourth

TABLE XXXI

TYPE II ERROR RATES-ALL USERS

	Errors		Attempts	Error Rate (%)	
	R*	NR**	R and NR	R	NR
Post Enrollment	U#	1220	19526	U	6.25
Normal	4779	3874	90410	5.[illegible]	4.28
Total	U	5094	109936	U	4.63

*R - with recycling
**NR- without recycling
#U = Unavailable. See Appendix C for a discussion of Type II errors with recycling.

phrase has been spoken. In Normal operation without recycling, a decision can be made on any phrase up to and including the fourth. From the definitions of recycling described here, it is expected that the Type II error rate will be higher with recycling than it is without. This did occur. As seen from Table XXXI, the Type II Error rates are much higher than the acceptable maximum rate of 2%.

In real time, when a phrase misregisters it (the same phrase) is reprompted later if needed in making the decision. When collecting impostor data, phrases were prompted until eight (two sets of four) registered phrases were collected and recorded. To do it exactly for Type II processing, each of the eight phrases would have had to have been registered and recorded three times (24 phrases) in order to have sufficient data for all possible playback cases. Rather than placing this burden on the impostors, recycling was simulated in data processing. The first four phrases were processed as usual, but if

a decision could not be made because of misregistered phrases, then a fifth (and sixth, if necessary in PE) phrase was used. If it registered, the decision function was computed and a decision was made as usual. If it misregistered, auto abort was invoked, and the last four phrases were used to make a decision. Once again, if a decision could not be made, the first (and second in PE) phrases were used and then the decision made. Thus, the difference between real time and playback is that in real time, the same phrase is recycled (reprompted) whereas, in playback, a new phrase was used.

For impostors, the probability of registering a phrase is important. In Phase II, 46.06% of all the phrases registered. In general, the conditional probability of a phrase registering, given that it has misregistered at least once, will be much smaller. Thus, the results without recycling are the most optimistic, and those with recycling are the most pessimistic. The actual results would probably be somewhere near a quarter of the way from the no recycling results to the recycling results.

The Type II Error rates for male vs. male and female vs. female both without and with recycling are shown in Tables XXXII and XXXIII, respectively. PE means that the victim,i.e., the reference file, is in PE. Normal means the victim is in the Normal verification state.

It is unlikely that any intruder would attempt to enter a secure area by claiming the identity of someone of the opposite sex. Thus, the monosexual male versus male and female versus female error rates are considered to be more important.

TABLE XXXII

TYPE II ERROR RATES WITHOUT RECYCLING

Category	Errors	Attempts	Error Rate (%)
Post Enrollment			
Male Vs. Male	1211	18846	6.43
Female Vs. Female	9	680	1.32
Normal			
Male Vs. Male	3801	86371	4.40
Female Vs. Female	73	4039	1.81

TABLE XXXIII

TYPE II ERROR RATES WITH RECYCLING

Category	Errors	Attempts	Error Rate (%)
Post Enrollment			
Male Vs. Male	U*	18846	U
Female Vs. Female	U	680	U
Normal			
Male Vs. Male	4666	86371	5.40
Female Vs. Female	113	4039	2.80

*U = Unavailable. See Appendix C for a discussion of Type II errors with recycling.

To determine if the ASV Type II error rates for male intruder against male reference file and for female intruder against female reference file are significantly different, the F ratio test is used.

In each of the following processing formats the value of

$$T = \frac{N_1}{N_2} \cdot \frac{e_2 + 1}{e_1 + 1} \; ; \; N_1, N_2 = \text{no. of female, male attempts respectively}$$

is greater than two:

1. PE processing with recycling;
2. Normal processing with recycling;
3. PE processing without recycling;
4. Normal processing without recycling.

The value of F(n,m) for each of these processing formats, is always less than two. Therefore, female against female Type II error rates are significantly different, at the 90% confidence level, than the male against male Type II errors for all of the processing formats. Only female versus female category, without recycling, meets the BISS requirement. These results along with the Type I results show that performance of the algorithm was shifted too far, especially for males.

5.2.5.2 Type II Errors and Speaker Averages Versus Entry Trials. The number of Type II Errors, the number of entry attempts, and the Type II Error rate are tabulated in Table XXXIV for the first 4,

the first 10, and first 40 trials where the results for those reference files with less than 4, 10 and 40 trials, respectively, have been removed. For the test period, 193 people had 4 or more trials, 176 people had 10 or more trials, and 45 people had 40 or more trials.

TABLE XXXIV

TYPE II ERRORS FOR THE FIRST N TRIALS AND REFERENCE FILES HAVING AT LEAST N TRIALS

Trials 1 through N	N=4	N=10	N=40
Errors	972	1824	634
Attempts	15449	35436	23042
Error Rate (%)	6.29	5.15	2.75

As noted earlier, PE (i.e., the first four successful verifications) use a different strategy than do subsequent verifications. In PE four registered phrases are required for a verification to occur. Thereafter, a decision can be made on any phrase from the first to as many as eight. This implies that it should be more difficult (lower error rate) for an impostor to pass during PE than Normal verification. However, the individual speaker averages tend to decrease with trials as seen in Figure 13 and from Table XXXIV it is clear that it becomes more difficult for an impostor to pass the more trials the individual has completed. The group with N=4 has virtually the same Type II error rate as all users in PE (6.29% versus 6.25%). The group with N=10 is higher than the total population in Normal (5.15% versus

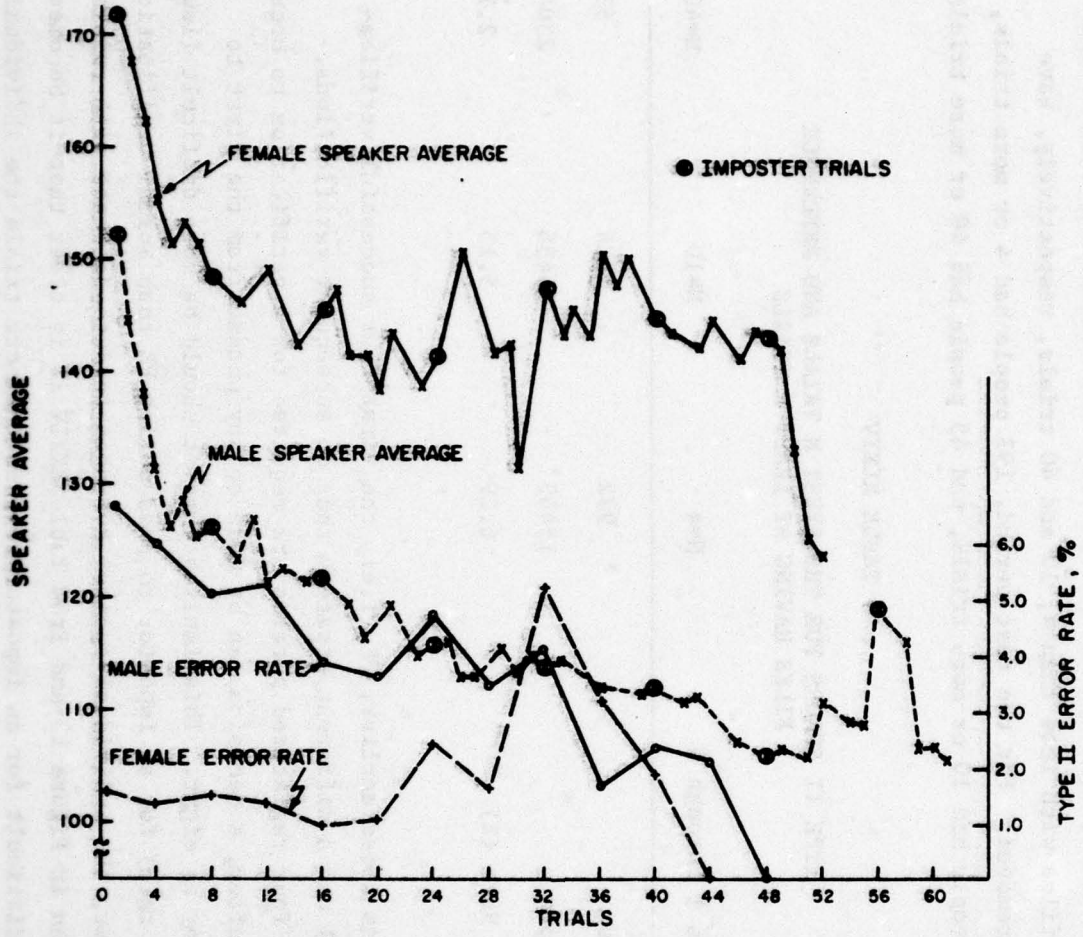


Figure 13 SPEAKER AVERAGES AND TYPE II ERROR RATES VERSUS TRIALS

4.28%). However, the subset with N=40 has an error rate much smaller than the 4.28%, i.e., 2.75%. This means that these 45 people had a much lower Type II error rate than did the other 131 people who also comprised the N=10 subset. The trend in the error rate, when the number of trials are near PE, depends on how fast the speaker averages decrease. If the speaker averages drop rapidly, one would expect that shortly after PE the error rate will decrease. This is what occurs here.

Figure 13 shows the variation of male and female speaker averages and Type II error rate as a function of the number of trials. The speaker averages drop off more rapidly during PE than in Normal operations. This is expected due to the larger weight factor that is used in updating during PE. For high trial numbers oscillations shown on the curve are due to the fact that fewer people are contributing to the data. The Type II error rates appear to follow the speaker average curves as expected. The error rates were plotted at every fourth trial. Figure 14 presents only the male Type II error rate versus trial. The error rates plotted are an average for groups of 5 trials. The 90% confidence limits (using chi-squared) are also indicated on the curve.

5.2.5.3 Type II Error Rate Versus Expected Scanning Error (ESE). Type II error rate versus expected scanning error is shown in Figure 15. This figure is a combination of male versus male and female versus female results but since there are only 7 females, it is essentially a male versus male result. From the use of the ESE, discussed in 5.2.2.3, 5.1.4 and 4.2.5.3 reference files with

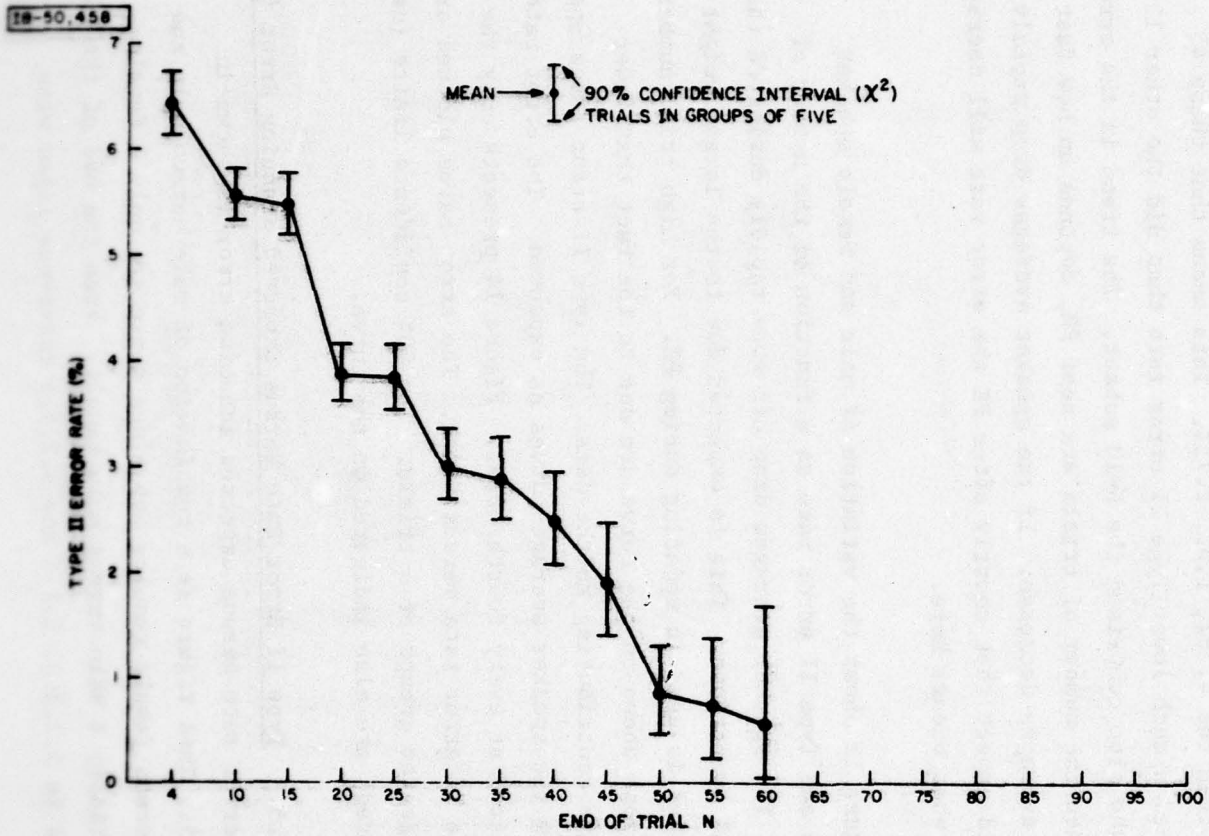


Figure 14 PHASE II TYPE II ERROR RATE VERSUS TRIAL

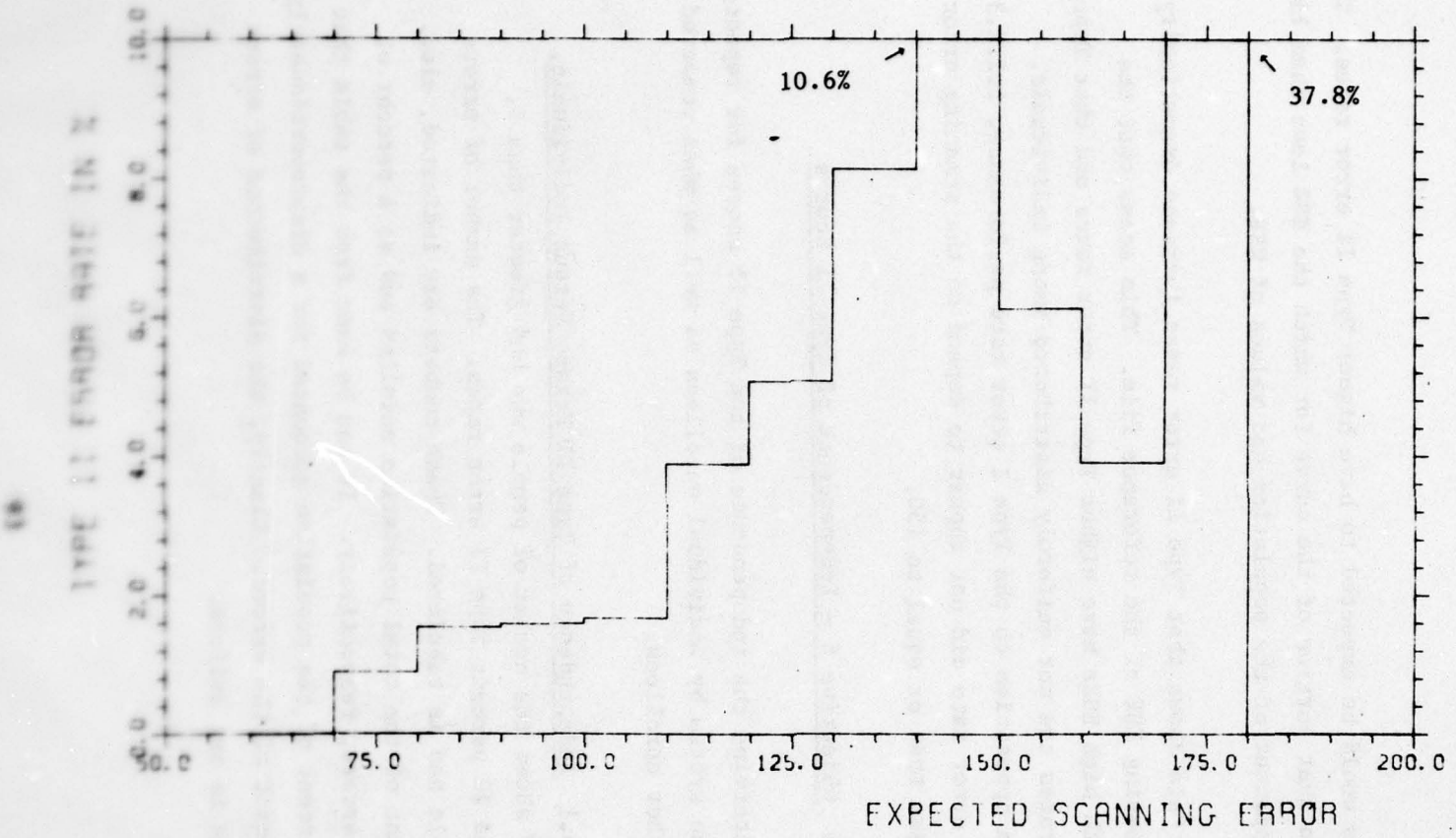


FIGURE 15 TYPE II ERROR RATE VERSUS EXPECTED SCANNING ERROR

AD-A056 772 MITRE CORP BEDFORD MASS F/G 17/2
TESTS RESULTS ADVANCED DEVELOPMENT MODELS OF BISS IDENTITY VERI--ETC(U)
JUL 78 M J FOODMAN F19628-77-C-0001
UNCLASSIFIED MTR-3442-VOL-2. ESD-TR-78-150-VOL-2 NL

2 OF 2
ADA
056772

END
DATE
FILMED
9-78
DDC

high ESE's would be expected to have higher Type II error rates. This is seen in that portion of the curve for which the ESE less than 150, where 90 percent of the population had values of ESE.

This data shows that Type II error rates increase dramatically with increasing ESE of the reference file. This means that the people with high ESEs have higher Type II error rates and that Type II error rates are not uniformly distributed among individuals. This is in opposition to the Type I error rate performance, 5.2.2.3, where the error rate did not appear to depend on the scanning error for ESE less than or equal to 150.

5.2.6 Objective 6 - Independence of Type II Scores

To determine the independence of the Type II scores for repeated uses of the system by individual enrollees as well as when compared against other enrollees.

5.2.6.1 Independence of Type II Errors Versus Individuals. Table XXXV shows the number of people who had greater than 5, 10, 15, and 20 percent Type II error rates. The number of errors these people had is tabulated. These numbers are indicated, also, as a percent of the total population enrolled and as a percent of the total errors, respectively. It can be seen from the table that a small percent of the population accounted for a disproportionately larger percent of the errors. Clearly, the distribution of errors among users is not uniform.

TABLE XXXV

PERCENT OF TOTAL POPULATION AND TOTAL ERRORS
AS FUNCTIONS OF THE TYPE II ERROR RATE

	Type II Error Rate			
	5%	10%	15%	20%
No. of People	53	27	16	5
% of Total Population Enrolled (200)	27%	14%	8%	2.5%
No. of Errors	4070	2980	2169	959
% of Total Errors (7834)	52.0%	38.0%	27.7%	12.2%

5.2.6.2 Distribution of Type II Decision Function Scores. A distribution of frequency of occurrence versus decision function score, for all users in PE and Normal, can be plotted for intruders who registered (see 4.2.6.1 for a discussion of registering) against a reference file. To compare the distribution of occurrence for the first phrase said to the distribution for succeeding phrases, each distribution is divided by the total number of times that the phrase registered. The normalized distributions of frequency of occurrence versus decision function score are plotted on top of the other for each phrase in Figure 16. These pdf's are exactly as used in the ASV algorithm as opposed to how they were used in Phase I.

The means of the distributions are 180.7, 180.2, 179.0, 176.3 for phrase 1 through 4, respectively. The standard deviations of

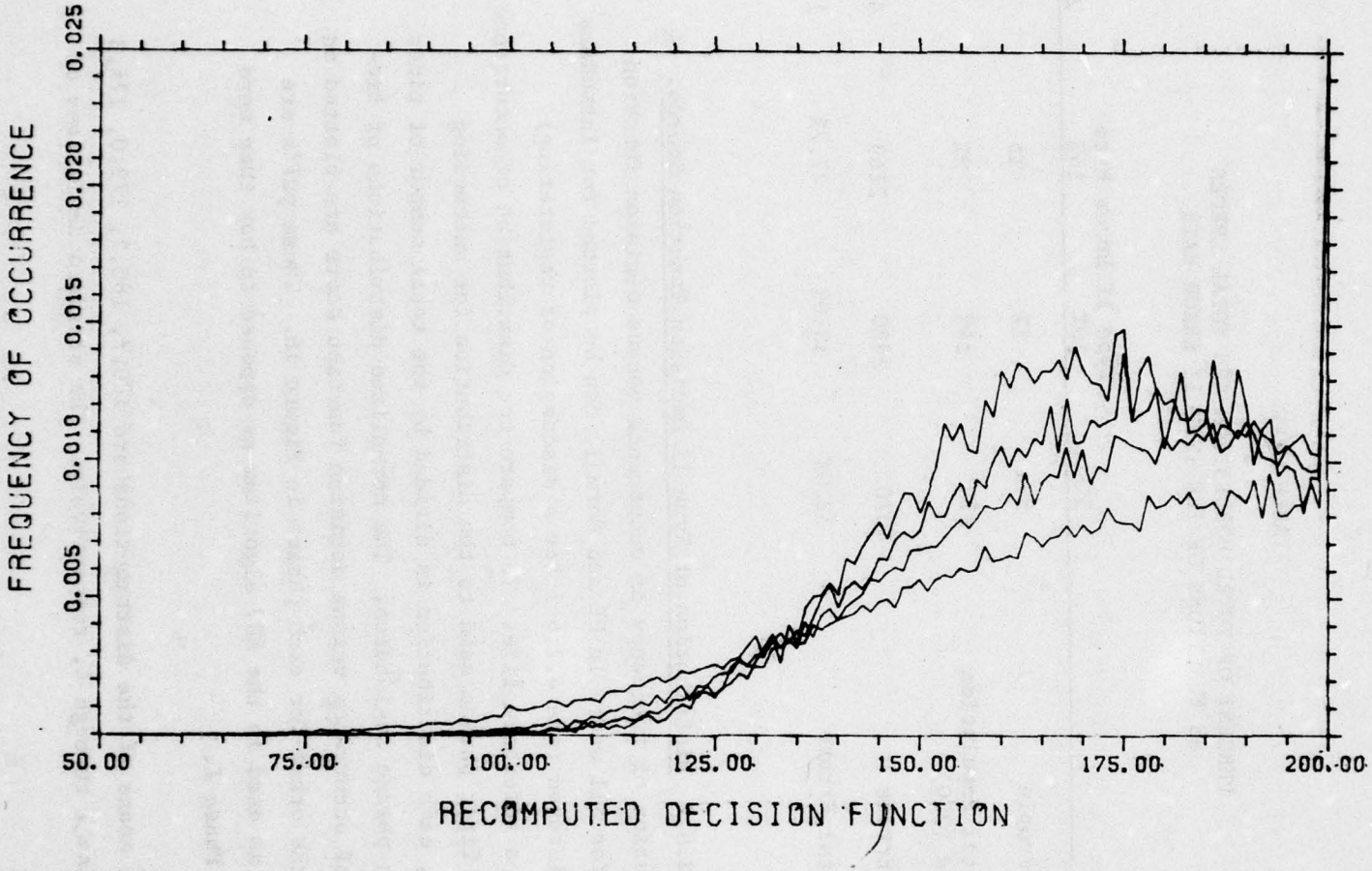


FIGURE 16 FREQUENCY OF OCCURRENCE VS. RECOMPUTED DECISION FUNCTION TYPE II

the distributions are 26.0, 23.1, 22.2, 21.8 for phrases 1 through 4 respectively. Figure 16 was derived from the data with limiting at 200, i.e., scores greater than 200 were assumed to be 200. For the data without limiting the means would either stay the same or increase and the standard deviation would stay the same. The variable denominator (see 5.1.4) would also contribute to lowering the mean. Thus, the decision function scores tend to become more concentrated nearer the mean for the later phrases. The occurrences shown here are not the total number of intruder attempts, but only those attempts in which the intruders speech pattern registered against the reference pattern.

5.2.7 Objective 7 - Sensitivity Analysis of Type I and Type II Errors to Thresholds

To determine the sensitivity of Type I and Type II Error variations to changing thresholds.

5.2.7.1 Sensitivity Analysis. This section shows the sensitivity of Type I and Type II errors to the value of the threshold against which the decision function is compared. Figure 17 was obtained by integrating the curves in Figures 12 and 16 and presenting the results for phrase one from each in Figure 17.

Figures 18, 19, and 20 for phrases 2, 3 and 4 were generated in an identical manner. The first curve in Figure 17 shows the fraction of all the Type I decision functions which were less than a particular decision function value. Type I decision functions are

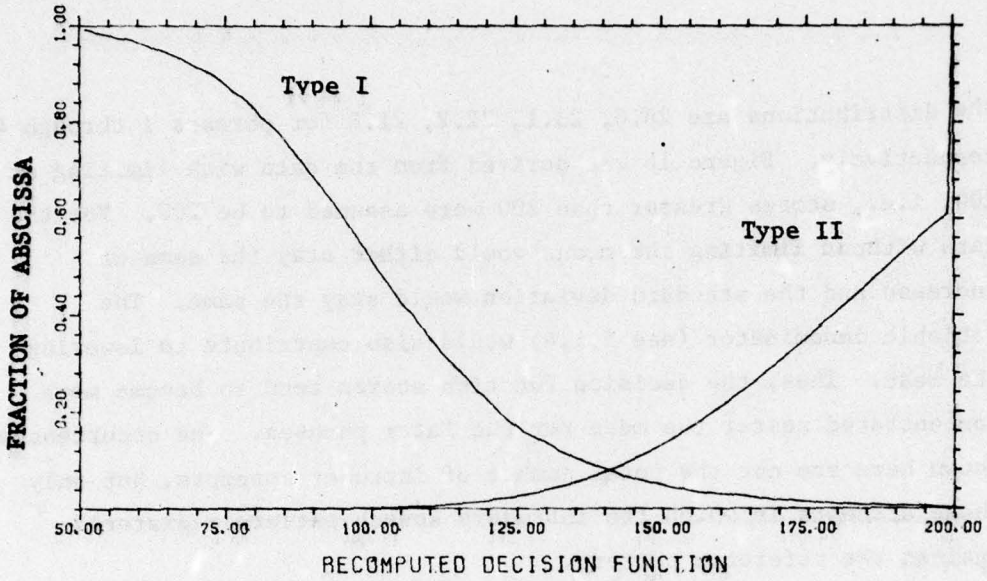


FIGURE 17 FRACTION OF TYPE I "RECOMPUTED DECISION FUNCTION SCORE" (RCDFS) GT AND TYPE II RCDFS LT ABSCISSA PHRASE 1

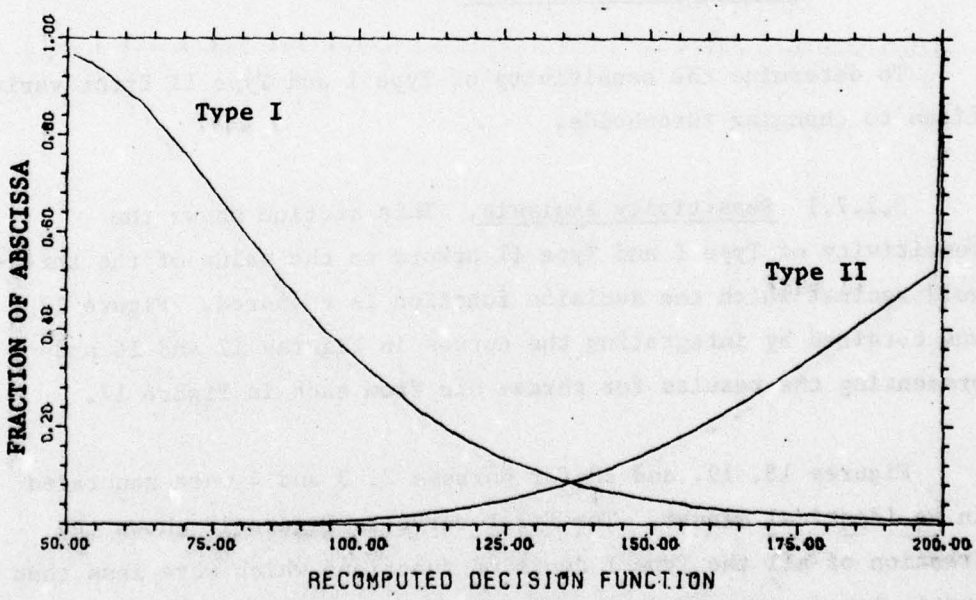


FIGURE 18 FRACTION OF TYPE I RCDFS GT AND TYPE II RCDFS LT ABSCISSA PHRASE 2

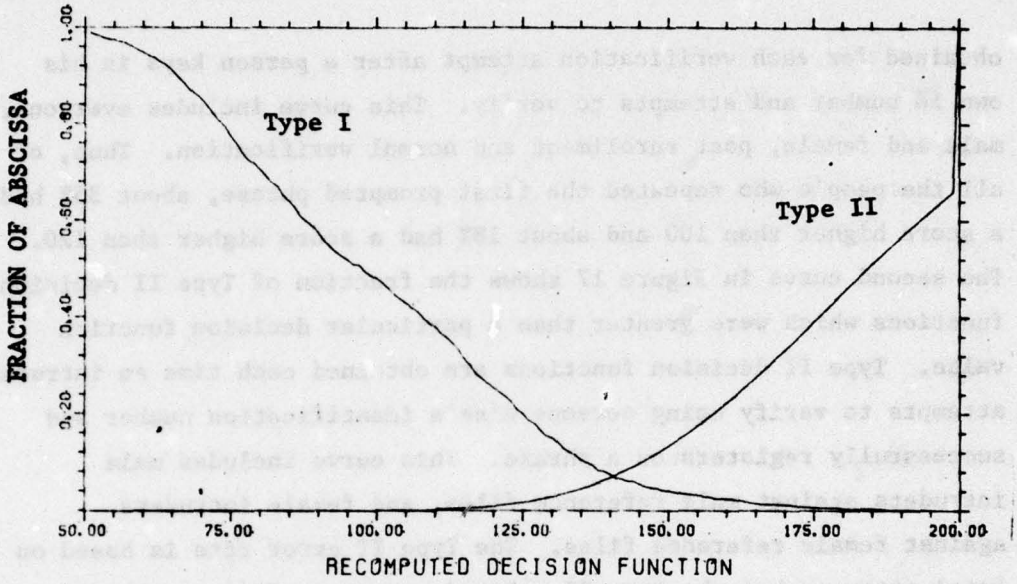


FIGURE 19 FRACTION OF TYPE I RCDFS GT AND TYPE II RCDFS LT ABSCISSA PHRASE 3

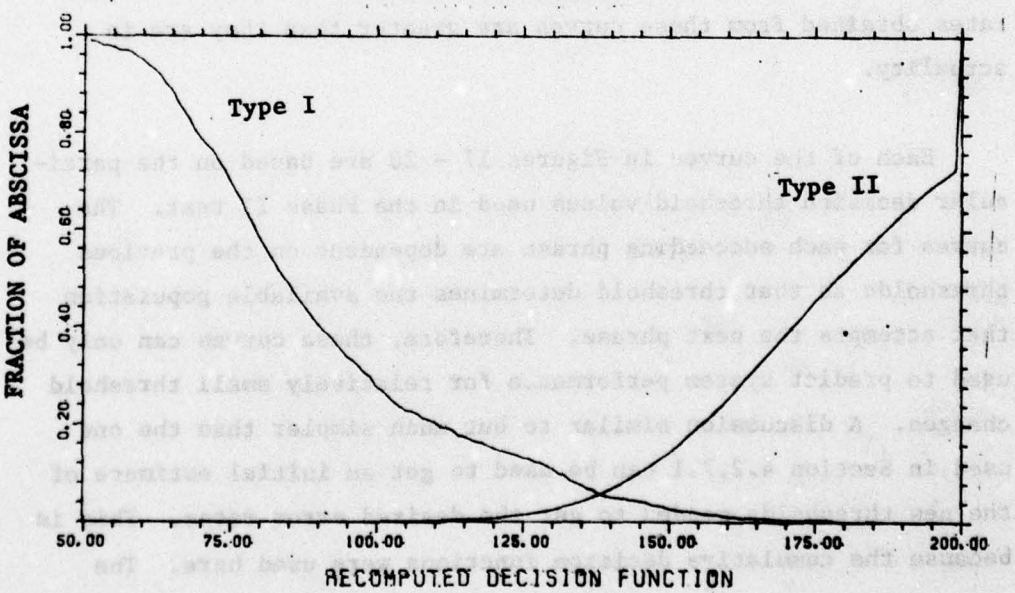


FIGURE 20 FRACTION OF TYPE I RCDFS GT AND TYPE II RCDFS LT ABSCISSA PHRASE 4

obtained for each verification attempt after a person keys in his own ID number and attempts to verify. This curve includes everyone, male and female, post enrollment and normal verification. Thus, of all the people who repeated the first prompted phrase, about 35% had a score higher than 100 and about 18% had a score higher than 120. The second curve in Figure 17 shows the fraction of Type II decision functions which were greater than a particular decision function value. Type II decision functions are obtained each time an intruder attempts to verify using someone else's identification number and successfully registers on a phrase. This curve includes male intruders against male reference files, and female intruders against female reference files. The Type II error rate is based on total attempts but the Type II related curves in Figures 17 - 20 are based on registered phrases only. Since the total attempts are greater than the number of registered phrases, the Type II error rates obtained from these curves are greater than they are in actuality.

Each of the curves in Figures 17 - 20 are based on the particular decision threshold values used in the Phase II test. The curves for each succeeding phrase are dependent on the previous thresholds as that threshold determines the available population that attempts the next phrase. Therefore, these curves can only be used to predict system performance for relatively small threshold changes. A discussion similar to but much simpler than the one used in Section 4.2.7.1 can be used to get an initial estimate of the new thresholds needed to get the desired error rates. This is because the cumulative decision functions were used here. The

actual error rates can then be determined by reprocessing the recorded data. This has not been done. The difference between total attempts and the number who registered a given number of phrases must also be taken into account.

5.2.8 Objective 8 - Verification Time Analysis

To determine the average time required for verification and the variance about this time.

5.2.8.1 Service Time (Verification Time). Service time is made up of keyboard time, verification time, first door opening-closing time, second door opening-closing time, and dead time. These last three times will not be discussed here. The average time to stroke four digits at the keyboard, hear the spoken digits and hit the SEND key was 3.36 seconds. The standard deviation about this time was 1.36 seconds. Table XXXVI shows a breakdown of the number of decisions, including those who did not verify, by phrase number. From Table XXXVI it was determined that an average of 4.15 phrases were required to verify during PE while during normal verification an average of only 1.53 phrases were required to verify. The average response time per phrase was 1.927 seconds. The standard deviation about this time was 0.34 seconds. Thus, the average verification time was computed (as described in 4.2.8.1) to be 15.9 seconds during PE but only 5.9 seconds during Normal verification. These times include the 1.9 seconds required to prompt a phrase.

Thus, after an individual has four successful verification attempts and is in Normal verification, the average time for keyboard and verification is 9.3 seconds. In an operational system only a small number of users would be new to the installation and be in PE

TABLE XXXVI

NUMBER OF DECISIONS VERSUS PHRASE NUMBER

Phrase Number	1	2	3	4	5	6	7	8	9	10
Post Enrollment	-	-	-	677	69	16	0	1	0	3
Normal	2741	1031	329	97	14	8	10	11	3	1

at any one time. A discussion of throughput is presented in Volume V, of this report.

5.3 PHASE II CONCLUSIONS

In the Normal mode of operation the Type I error was much less than the maximum acceptable rate of 1% but this occurred at the expense of a Type II error rate which was much greater than the maximum acceptable rate of 2%. The changes made to the Phase I ASV algorithm overcompensated and resulted in the high Type II error rate and low Type I error rate. Most of this was noted in the male performance.

5.3.1 Type I Error Rates Versus Decision Threshold

An attempt was made to determine the effect on Type I error rate of different values of decision threshold.

Table XXXVII shows Type I Error rates versus decision threshold. The small number of attempts was a result of using only impostor attempts. This was necessary as only impostor attempts, which force the entrant to say eight phrases, provide enough spoken phrases to

enable a comparison of error rates versus threshold. The old thresholds were 100, 120, 135, 145 on phrase 1 through 4, respectively and the new thresholds were 100, 120, 125, and 135. As seen from the table the changes in the threshold had no effect on the error rate. It was not practical to do similar processing for Type II data, but the effect of these new thresholds could be estimated from Figures 17 through 20.

TABLE XXXVII

TYPE I ERROR RATES VERSUS DECISION THRESHOLD

	New Thresholds			Old Thresholds		
	Errors	Attempts	Error Rate (%)	Errors	Attempts	Error Rate (%)
Males	0	742	0	0	742	0
Females	2	166	1.20	2	166	1.20
Total	2	908	0.22	2	908	0.22

5.3.2 A Comparative Discussion of Phase II Versus Phase I

The effect of each of the changes to the ASV algorithm presented in 5.1 is in order. They will be treated individually and in combination, as necessary. The primary purpose of the changes was to lower Type I error rate and increase the Type II error rate in PE. A secondary purpose was to lower the error rates in Normal as well.

Normalization by the standard deviation rather than the mean (5.1.2) should improve the signal-to-noise ratio and improve the definition of formant frequencies in the speech data which should

allow for more registered phrases, and other things being equal, a lower Type I error rate and a higher Type II error rate. The effect of this change on the observed results is relatively minor.

The change in the estimate of ESE at the end of enrollment to account for intersession variance (5.1.3) had a major impact on the results. In particular, the use of 140 as a minimum allowable ESE for each word affected almost every male enrollee. This is because the ESE for males is about 20% lower than it is for females. The intent of the constant was to aid those enrollees who were unusually stable during enrollment (a very small intrasession variance). A value of 110, or so, instead of 140 should have been used.

The change in the coefficient b from 1.17 to 1.25 (5.1.3) appeared to be necessary because the ESE increased for the first three sessions after enrollment, thus indicating the need for a larger coefficient. As one uses the ASV system regularly, one should expect the ESE to decrease. It did in Phase II, but the initial value was too high.

The third change (5.1.4), which was in the computation, also had its predominent effect on PE. The value of max_i was 140 in Phase I and 160 in Phase II. This, in combination with the change in 5.1.3, served to reduce Type I errors at the expense of many more Type II errors. A lower value, 140 or 145, is probably more in line with desired performance. A changing max_i with the number of phrases registered may or may not be necessary. Relative to Phase I, the Type I error rate should be higher on phrase 1 and 2, the same on phrase 3 and lower on phrase 4. Since about 85% of the users verify

on the first two phrases, the net effect on Type I errors and verification time should be minimal. Type II error rate will be smaller on phrases 1 and 2, the same on phrase 3 and higher on phrase 4 relative to Phase I. Since the probability of four contiguous registered phrases went from about 4% to about 16%, the fourth phrase will add additional Type II errors. The recycling of phrases in Normal (5.1.5) will also increase the Type II errors on phrase 4 slightly.

6.0 FIELD TEST

6.1 DESCRIPTION

The Field Test was conducted at Pease Air Force Base in New Hampshire from November 1976 thru February 1977. A detailed description of the overall system setup is provided in Volume V of this report.

The test population consisted of 274 people (267 males and 7 females) enrolled but only 207 people (201 males and 6 females) were recorded as using the system following enrollment. However, there were 5 males all with only one attempt and one error, whose errors were caused by wrong ID (4) and harassment (1). As indicated in the following section these errors are not considered. Therefore, there were in effect only 202 people (196 males and 6 females) who used the system following enrollment. Of the total population enrolled, 74% returned for at least one verification. The user population consisted of officers, non-commissioned officers, airmen, and civilians. Their jobs included mechanics, security police, administrators, electricians and general maintenance workers.

The ASV algorithm used in the Field Test was the same as the Phase I ASV algorithm with one modification: The Field Test allowed recycling (defined in 5.2.4.1) in Normal verification.

6.2 RESULTS

6.2.1 Objective 1 - Type I Error Analysis In Real Time

To estimate the Type I Error rate (failure to verify proper identity when the representation is actually true) and to determine the confidence limits of the Type I Error estimate.

6.2.1.1 Type I Errors. Table XXXVIII shows the total number of Type I errors which occurred in both PE and Normal processing. The errors were categorized by assignable cause where possible. There were, in addition, 10 intentional Type I errors, which have been removed from the data and are not considered in Table XXXVIII and any of the analysis that follows except where indicated otherwise.

The ID entry, for the ASV part of the entry control system, is made by keystroking the proper ID number. Entrants may forget their ID number or incorrectly keypunch their ID numbers resulting in a Type I error. In the Phase I and II tests, the ID number was echoed to the user as it was keystroked. In the Field Test it was not. Therefore, more incorrect ID numbers are likely. If the ID entry was made via a card reader then the number of wrong ID entries should be eliminated. Since the primary purpose of this test is to test the algorithm, the errors caused by entering the wrong ID and harassment of the entrant are also deleted from the statistics and are not considered in any of the analysis that follows except where indicated otherwise. One person who had false teeth accounted for 18 errors. This person was re-enrolled 3 times after having error rates of 35, 67, and 30 percent and a total of 13 errors. The false

TABLE XXXVIII

TYPE I ERROR RATES - ALL USERS

	Errors	Attempts	Error Rate (%)	Assignable Causes					
				Speech Errors	Colds	Wrong ID	Harass-ment	Recycling	Other
Post Enrollment	73	740	9.86	30	5	5	2	0	31
Normal	68	6430	1.06	6	15	7	4	9	27
Total	141	7170	1.97	36	20	12	6	9	58

teeth apparently interfered with his speech and his data prior to his last re-enrollment is not considered. Three other people were re-enrolled. The total of four people (1.98% of the total users) who were re-enrolled accounted for 20 errors in 54 attempts for an error rate of 37.04%. After re-enrollment the four people had a combined error rate of 6/294 = 2.04%. The person with the false teeth had 5 of the 6 errors. The number of users with false teeth who did not have difficulty verifying is not known. There were 9 successful Normal verifications which occurred after the eighth phase. These verifications would not have occurred if recycling was not used as in Phase I. These 9 successful verifications are considered as Type I errors in Table XXXVIII. Table XXXIX shows the error rates when wrong ID, pre-reenrollment, harassment, errors, and recycling errors are not considered.

TABLE XXXIX

TYPE I ERROR RATES EXCLUDING WRONG ID, PRE-REENROLLMENT, HARASSMENT ERRORS AND RECYCLING

	Errors	Attempts	Error Rate (%)
Post Enrollment	54	703	7.68
Upper Bound*			9.14
Normal	40	6395	0.63
Upper Bound*			0.77
Total	94	7098	1.32

*90% confident that the true error rate is less than the upper bound. See Appendix B.

Type I errors resulting from speech related causes accounted for 41% of PE errors but for only 10% of the Normal errors. Speech related errors include mispronouncing a word or words, forgetting one or more of the prompted words and difficulty saying the prompted word. Most of these errors occurred during PE when people were least familiar with the system operation. To determine if the PE error rate is significantly different from the Normal error rate the F ratio test is used:

$$T = \frac{6395}{703} \cdot \frac{55}{41} = 12.2$$

$$F_{p=.9}(82,110) = F(82,110) = 1.34$$

Since F is less than T, the error rates are significantly different at the 90% confidence level.

There is the possibility that, due to the initial nervousness of new users (as seen here by the large number of speech errors), the PE error rate will always be higher than the Normal error rate. The combined PE and Normal, male and female Type I error rate is 1.32%. The fact that this error rate is greater than 1% is due mainly to the very high error rate (7.68%) that occurs, as expected, during PE.

6.2.2 Objective 2 - Type I Error Analysis In Non-Real Time

To determine how various parameters affect the Type I Errors and system performance.

6.2.2.1 Type I Errors Versus Sex. Table XL tabulates the Type I error rates versus sex. Set 1 contains all errors while Set 2 has the assignable causes indicated in Table XXXIX deleted. This data shows the male and female error rates during PE are more than 10 times as high as in Normal verification. To determine if male and female error rates in PE as shown in Set 2 are significantly different the F ratio test is used:

$$T = \frac{676}{27} \cdot \frac{4}{52} = 1.93$$

$$F(104,8) = 2.28$$

Since F is greater than T, it cannot be said, at the 90% confidence level, that the error rates are significantly different.

TABLE XL

TYPE I ERROR RATES VERSUS SEX

Set 1

	Errors	Attempts	Error Rate (%)
Male P.E.	70	713	9.82
Male Normal	65	5974	1.09
Female P.E.	3	27	11.11
Female Normal	3	456	0.66

Set 2

	Errors	Attempts	Error Rate (%)
Male P.E.	51	676	7.54
Male Normal	38	5940	0.64
Female P.E.	3	27	11.11
Female Normal	2	455	0.44

For Normal verification the F test yields:

$$T = \frac{5940}{455} \cdot \frac{3}{39} = 1.00$$

$$F(78,6) = 2.75$$

Again, it cannot be said, at the 90% confidence level, that the female error rate is significantly different than the male error rate.

Because only six females participated in this test, a general conclusion cannot be made.

6.2.2.2 Type I Errors Versus Time of Day. Type I errors versus time of day is shown in Table XLI. Applying the F test to PE, Normal and total data for morning and afternoon yields:

$$T_P = \frac{394}{309} \cdot \frac{27}{29} = 1.19; \quad F(58,54) = 1.41$$

$$T_N = \frac{2801}{3594} \cdot \frac{27}{15} = 1.40; \quad F(30,54) = 1.51$$

$$T_T = \frac{3110}{3988} \cdot \frac{55}{41} = 1.05; \quad F(82, 100) = 1.34$$

Therefore, it cannot be said that the error rates in the morning and afternoon are significantly different.

TABLE XLI

TYPE I ERRORS VERSUS TIME OF DAY

	Errors	Attempts	Error Rate (%)	Upper Bound (%)
Morning				
Post Enrollment	28	394	7.10	9.14
Normal	26	3594	0.72	0.90
Total	54	3988	1.35	1.51
Afternoon				
Post Enrollment	26	309	8.41	11.00
Normal	14	2801	0.50	0.72
Total	40	3110	1.29	1.75

6.2.2.3 Type I Error Rates Versus Expected Scanning Error*. The expected scanning error (ESE) is a measure of the consistency between repetitions of the same words. The lower the ESE, the better the match between reference patterns and new speech material. The ESE is used to normalize the current scanning error and to determine a decision function which is compared against a threshold.

Figure 21 shows the Type I error rate versus ESE for the combined data of males and females in both Normal and PE. This figure was derived from the value of ESE that each individual had at the end of the test. For each 10 units of ESE the Type I error rate was calculated by summing the individual errors and dividing by the sum of the individuals attempts. The Type I error rate increased for those people with an ESE greater than 140, as expected.

*Expected Scanning Error and speaker average refer to the same quantity and are used interchangeability in this report.

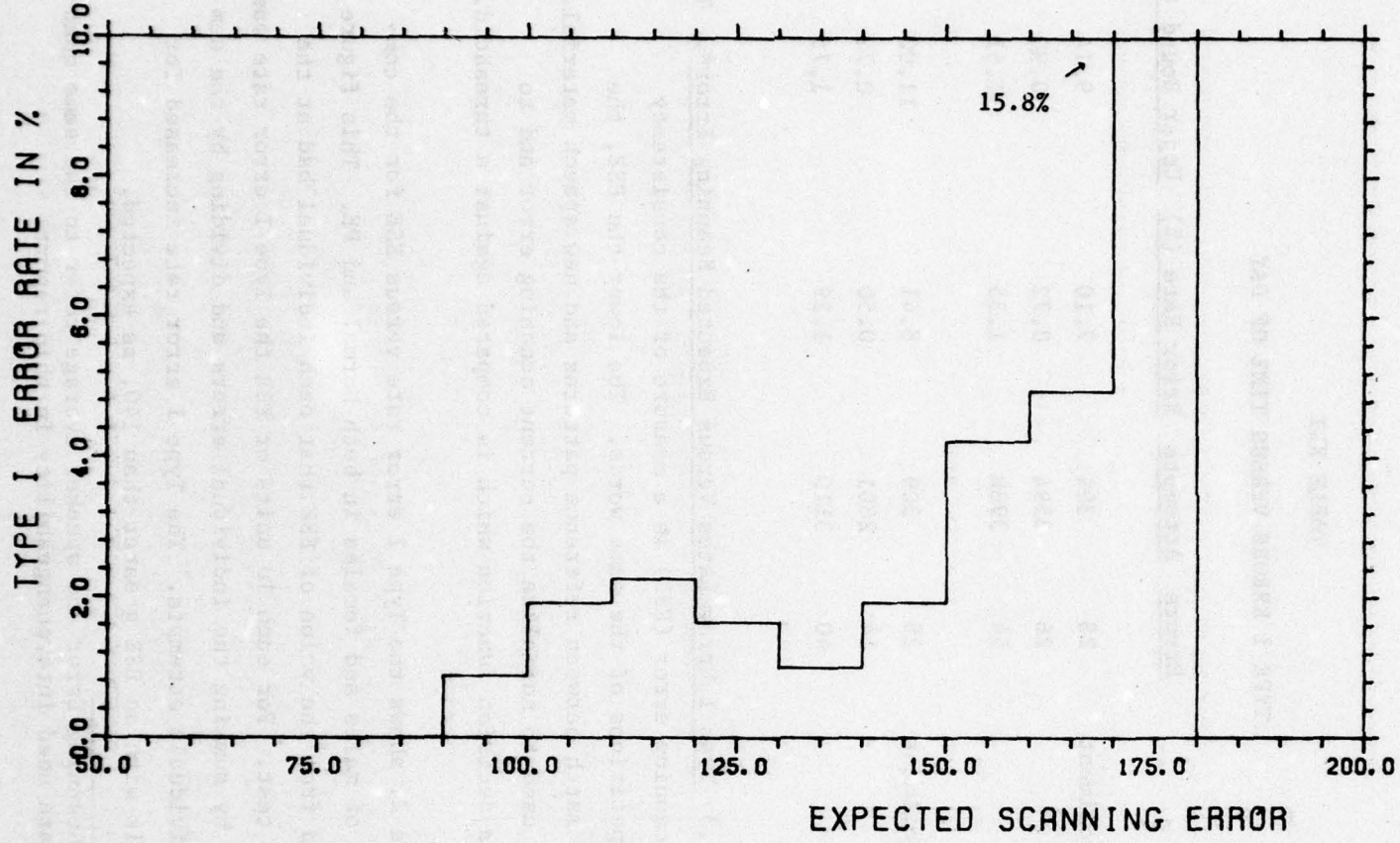


FIGURE 21 TYPE I ERROR RATE VERSUS EXPECTED SCANNING ERROR

6.2.2.4 Type I Errors Versus Station. This section does not apply since only one station was used in this test.

6.2.2.5 Type I Errors Versus Entry Trials Since Enrollment. Of the 274 people enrolled, 99 males and 6 females had ten or more trials. Because of the low number of females in this category only the combined male and female errors are considered. The Type I errors versus trials for these people are shown in Table XLII. As seen from Table XLII the error rate found in the first four trials, PE, is greater than any other group of four trials thereafter.

TABLE XLII

TYPE I ERRORS FOR USERS WITH 10 OR MORE TRIALS

Trial Number	Errors	Error Rate (%)
1	11	10.50
2	6	5.71
3	3	2.85
4	3	2.85
5	5	4.76
6	1	0.95
7	1	0.95
8	1	0.95
9	1	0.95
10	2	1.90

After the first five trials, there are so few Type I errors that no variation in Type I errors versus increasing trial number can be determined. The high error rate at trial 5 is due in part to the affect of residual PE at this trial. Therefore, more data is required to determine the correlation, if any, between trial number and Type I errors.

6.2.2.6 Type I Errors Versus Phrases Required To Enroll. Type I errors versus the number of phrases required during the person's last enrollment are tabulated in Table XLIII. This is the combined PE and Normal data.

For any category of phrases greater than 26, there is no more than two errors and there are entry attempts from no more than two persons. This is not enough data to be statistically significant. Therefore, the data was combined into two categories as in Phase I; less than or equal to 21 phrase required and greater than 21 phrases required. These categories are shown at the bottom of Table XLIII. Applying the F test to these two groups yield:

$$T = \frac{4700}{2398} \cdot \frac{43}{53} = 1.59$$

$$F(106,86) = 1.34$$

Therefore, as a group, people requiring more than 21 phrases to enroll did have significantly different error rates than those requiring 20 or 21 phrases to enroll.

TABLE XLIII

TYPE I ERROR RATE VERSUS PHRASES REQUIRED
DURING ENROLLMENT

Phrases Required	Errors	Attempts	Error Rate (%)
20	33	3301	1.00
21	19	1399	1.36
22	5	543	0.92
23	14	374	3.74
24	2	157	1.27
25	4	323	1.24
26	12	554	2.17
27	2	7	28.57
28	0	120	0
29	0	0	0
30	0	188	0
31	0	98	0
32	0	0	0
33	1	0	0
34	0	0	0
35	1	2	50.00
36	1	32	3.12
$\leq$ 21	52	4700	1.10
$\geq$ 22	42	2398	1.75

6.2.2.7 Type I Error Rate Versus Day Of Week. Type I error rate versus the different days of the week as shown in Table XLIV indicates a decreasing error rate as the week progresses. There were three extended weekends during the test. A consecutive Monday, Tuesday and Wednesday during the test was a holiday and during that week, Thursday was the first work day of the week. If the data from that Thursday is included in the Monday column, the results are as shown in Table XLIV. These results include all the data in Table XXXVIII plus the 10 intentional errors.

TABLE XLIV

TYPE I ERROR RATE VERSUS DAY OF WEEK

	Mon	Tues	Wed	Thurs	Fri
Normal Week Error Rate (%)	3.04	2.27	1.44	1.56	0.90
Holiday Week Error Rate (%)	3.10	2.27	1.44	1.41	0.90

6.2.2.8 Type I Errors Versus Personal Statistics. Tables XLV through XLVIII tabulate the Type I errors as functions of the entrant's height, education level, age, and primary education location. However, there are an insufficient number of users in each category so that no reliable trend has been developed.

TABLE XLV

TYPE I ERRORS VERSUS HEIGHT

Height (h)	Errors	Attempts	Error Rate (%)
h <61	1	110	0.9
61 ≤h <64	4	160	2.50
64 ≤h <67	10	795	1.26
67 ≤h <70	20	2920	0.68
70 ≤h <73	48	2291	2.10
73 ≤h	11	822	1.33

TABLE XLVI

TYPE I ERRORS VERSUS EDUCATION LEVEL

Years of School	Errors	Attempts	Error Rate (%)
≤ 8	0	59	0
< 12	4	45	8.89
= 12	35	4384	0.80
< 16	50	2067	2.42
≥ 16	5	543	0.92

TABLE XLVII

TYPE I ERRORS VERSUS AGE

Age (Yrs.)	Errors	Attempts	Error Rate (%)
≤25	50	4158	1.20
26-35	15	1687	0.89
36-45	17	1157	1.47
46-55	10	92	10.87
>55	2	4	50.00

TABLE XLVIII

TYPE I ERRORS VERSUS PRIMARY EDUCATION LOCATION

Location	Errors	Attempts	Error Rate (%)
New England	34	2566	1.33
New York Area	28	2481	1.13
South	19	1024	1.86
West	11	767	1.43
Canada	2	198	1.01
Scandinavia	0	62	0

6.2.3 Objective 3 - Independence Of Type I Scores

To determine the independence of the Type I scores for repeated uses of the system by individual enrollees as well as when compared against other enrollees.

6.2.3.1 Independence Of Type I Errors Versus Individuals. Of the 274 people enrolled in the ASV system 202 had one or more entry attempts. Most had no Type I errors at all. Table XLIX shows the number of people who had the indicated number of errors.

TABLE XLIX

TYPE I ERRORS VERSUS INDIVIDUALS

Type I Errors (N)	Number of People With N Type I Errors	Total Errors
0	152	0
1	30	30
2	9	18
3	5	15
4	2	8
5	2	10
6	1	6
7	1	7

A minority of those using the system (24.7%) accounted for all the Type I errors. Five percent of all the people using the system accounted for almost half (49%) of the Type I errors. Thus, the Type I errors are not uniformly distributed among all users of the system.

Table L shows the number of people who had greater than 3, 5, and 10% Type I error rates. The number of errors these people had is tabulated. These numbers are indicated, also, as a percent of the total population (202 who had at least one trial) and as a percent of the total errors, respectively. As seen from the table, 17.8% of the people had 80% of the errors. For those with more than one error and from 6.2.2.3, there appears to be a non-uniform distribution of errors among users.

TABLE L

PERCENT OF TOTAL POPULATION AND TOTAL ERRORS AS FUNCTIONS OF THE TYPE I ERROR RATE

	Type I Error Rate		
	3%	5%	10%
No. of People	36	29	23
% of Total Population (202 who had 1 trial)	17.8%	14.4%	11.4%
No. of Errors	75	62	33
% of Total Errors (94)	80%	66%	35%

6.2.3.2 Distribution Of Type I Decision Function Scores. To compare the distribution of occurrence versus decision function score for the first, second, third, and fourth phrases, the number of occurrences of a given decision function score for a given phrase is divided by the number of times that phrase is used. This yields a probability density function (pdf). The resulting four pdf's are plotted in Figure 22. These plots are for all users in PE and Normal. The means of the distributions are 98.9, 114.6, 122.0, and 119.8 for phrases 1 through 4, respectively. The standard deviation of the distributions are 30.2, 24.1, 25.0 and 29.4 for phrases 1 through 4, respectively.

The functions have means which tend to increase with increasing phrase number. This is because a cumulative decision strategy as described in 4.2.3.2 was used. The cumulative decision strategy used past values of the scanning error and expected scanning error to calculate the decision function scores. The decision function is used to determine verification using a different threshold after each phrase.

6.2.4 Objective 4 - Type II Error Analysis In Real Time

To estimate the Type II Error rate (erroneously verifying identity when the representation is actually false) and to determine the confidence limits of the Type II Error when mimics attempt verification.

No mimic tests were run during the Field Test. Tests against this algorithm are discussed in 4.2.4.

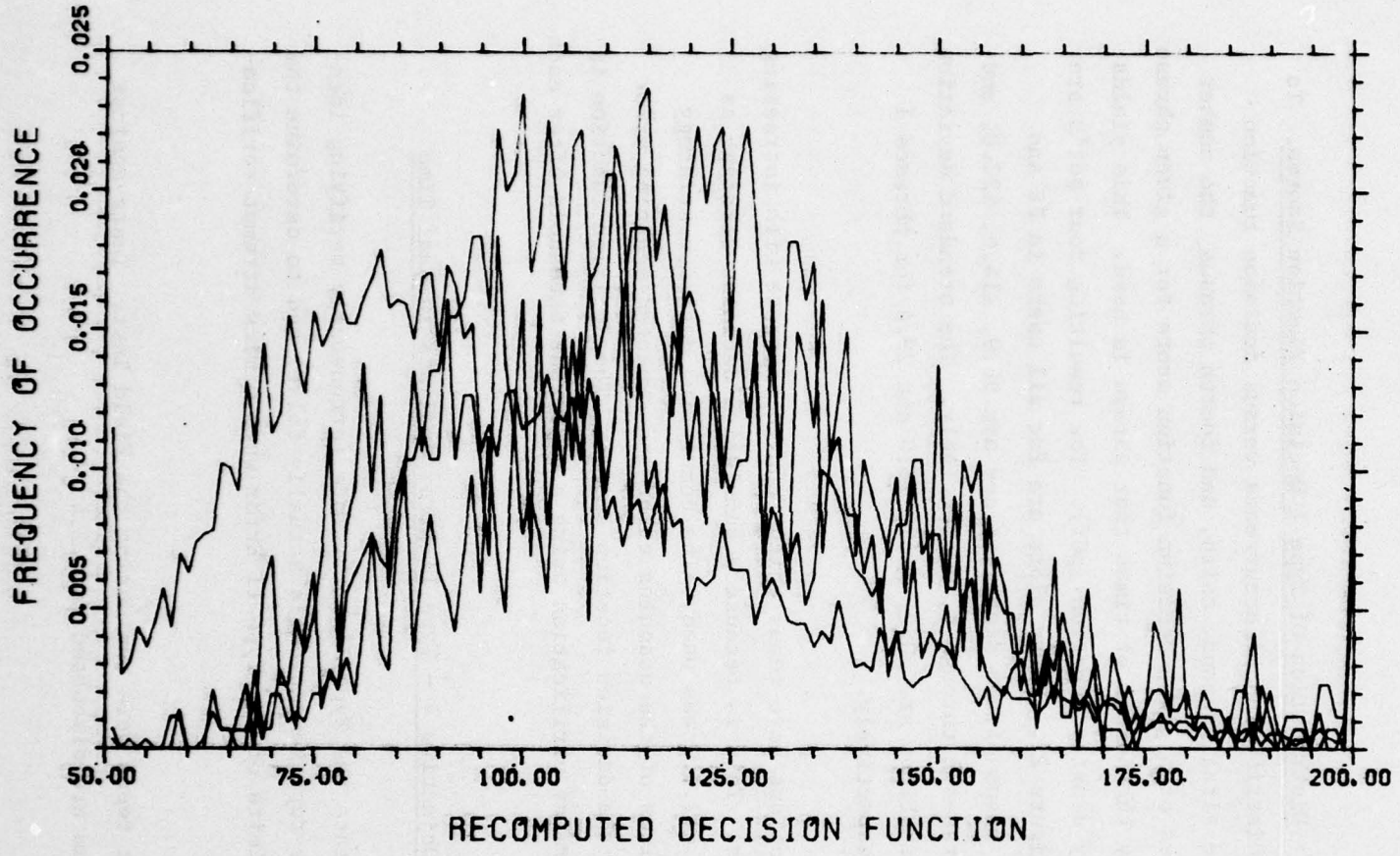


FIGURE 22 FREQUENCY OF OCCURRENCE VS. RECOMPUTED DECISION FUNCTION TYPE I

6.2.5 Objective 5 - Type II Error Analysis In Non-Real Time

To estimate the Type II Error (erroneously verifying identity when the representation is actually false) and to determine the confidence limits of the Type II Error when matching within the enrolled population is performed.

6.2.5.1 Type II Errors. The Type II error testing was conducted as described in 4.2.5.1. The Type II errors considered in the following sections are the sum of the male vs. male and female vs. female errors. The number of females are too small to consider these two groups separately. Table LI shows the total number of Type II errors which occurred in both PE and Normal processing.

TABLE LI

TYPE II ERROR RATES-ALL USERS

	Errors		Attempts	Error Rate (%)	
	R*	NR**	R and NR	R*	NR**
Post Enrollment	U#	262	36904	U	0.71
Normal	1124	864	26411	4.26	3.27
Total	U	1126	63315	U	1.78

*R results using recycling
**NR results not using recycling
#U = Unavailable. See Appendix C for a discussion of Type II errors with recycling.

When recycling is not used the Type II error rate for the combined PE and Normal data is less than the maximum acceptable value of 2% whereas with recycling it is not. However, the Type II error rate for Normal Verification is more important for use of the system over an extended period. (Recycling was defined in 5.2.5.1.) It is expected that the Type II error rate will be higher with recycling than it is without recycling. This is due to the recycling condition of allowing repetition of misregistered phrases if required. The effect is greater in PE than Normal since two phrases are recycled rather than one.

An explanation of the higher than desired error rate in Normal verification is provided in 7.2.

6.2.5.2 Type II Errors And Speaker Averages Versus Entry Trials. The number of Type II errors, the number of entry attempts, and the Type II error rate are tabulated in Table LII for the first N trials, where those reference files with less than N = 4, 10, 40, 60 and 80 trials, respectively, have been removed. 138 people had 4 or more trials, 104 people had 10 or more trials, 81 people had 40 or more trials, 67 people had 60 or more trials, and 35 people had 80 or more trials.

As noted earlier, the first four successful verifications use a different strategy than do subsequent verifications. In PE, four registered phrases are required for a verification to occur. Therefore, the fact that it is more difficult to pass on someone else's ID during PE than in the beginning portion of Normal verification is not unreasonable. Thereafter, a decision can be made on any phrase from the

TABLE LII

TYPE II ERRORS FOR THE FIRST N TRIALS AND FILES HAVING AT LEAST N TRIALS

Trials 1 Through N	N = 4	N = 10	N = 40	N = 60	N = 80
Errors	106	89	316	353	257
Attempts	9503	6366	11,223	12,937	7657
Error Rate (%)	1.12	1.40	2.82	2.73	3.36

first to as many as 8 without recycling and 10 with recycling. As per the discussion in 5.2.5.2, the slow decline in the value of the speaker averages with trials in the Field Test indicate that a slow decline in the error rates as trials increase should be expected. The expected decline of the Type II error rate in Normal verification with increasing trials did not occur. It is possible that a few of the people with a large number of trials had similar reference patterns and as the trials increase these people tend to dominate the results. This could account for the increasing error rate with trial.

Figure 23 shows the variation of the combined female and male speaker averages, Type I error, and Type II errors as functions of the number of trials. The speaker average is plotted for every eighth trial. A closer look at PE shows that the speaker average actually increases to a maximum of 131 at trial 4 and decreases thereafter. Oscillations in the curves are due to the small amount of data at the large values of trial number. The Type II error rate, which was calculated at every fourth trial point, does not appear to follow the

IA-50,453

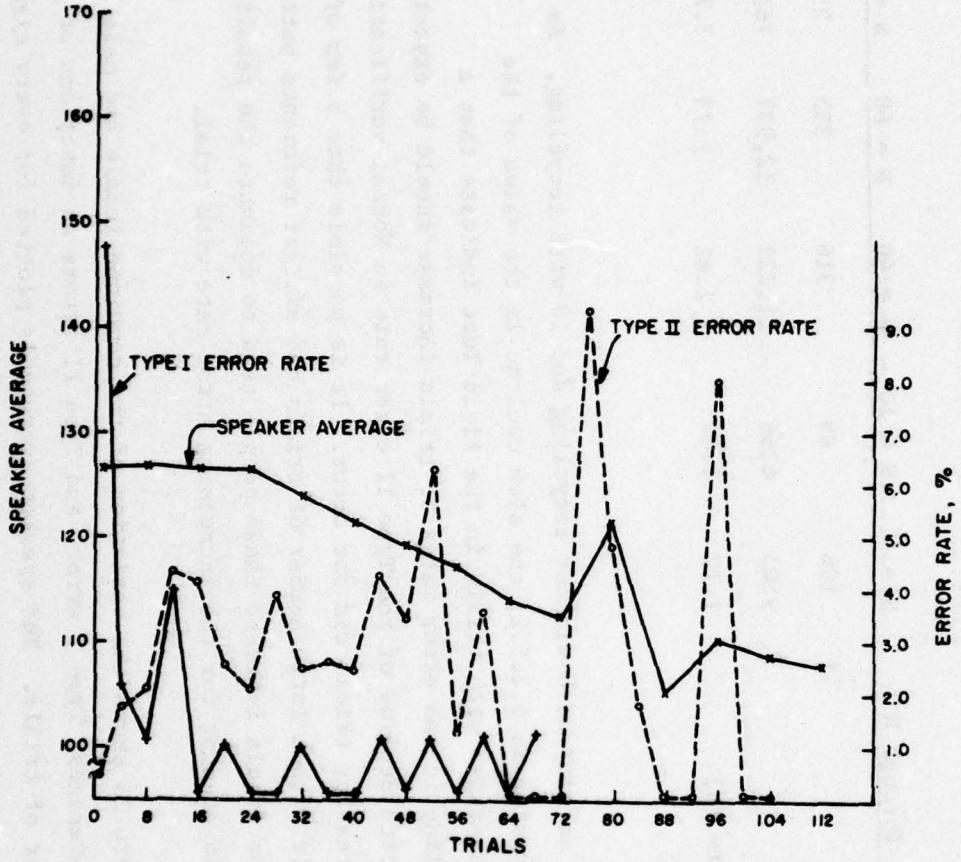


Figure 23 ERROR RATE AND MALE SPEAKER AVERAGES VERSUS TRIALS

speaker average contrary to what was expected. This indicates that the error rate depended strongly on who happened to be the imposters on a given trial number.

6.2.5.3 Type II Error Rate Versus Expected Scanning Error (ESE). Type II error rate versus expected scanning error is shown in Figure 24. This figure is a combination of male versus male and female versus female results. Because of the much larger male population this figure basically reflects male versus male results. From the use of the ESE, discussed in 6.2.2.3, reference files with high ESE's would be expected to have higher Type II error rates. This is seen in that portion of the curve for which the ESE is less than 150, where 87 percent of the population had values of ESE.

This data shows that Type II error rates increase dramatically with increasing ESE of the reference file. This means that the people with high ESE have higher Type II error rates and that Type II error rates are not uniformly distributed among individuals. This is in opposition to the Type I error rate performance, 6.2.2.3, where the Type I error rate did not appear to depend on the ESE for ESE values less than or equal to 150.

6.2.6 Objective 6 - Independence Of Type II Scores

To determine the independence of the Type II scores for repeated uses of the system by individual enrollees as well as when compared against other enrollees.

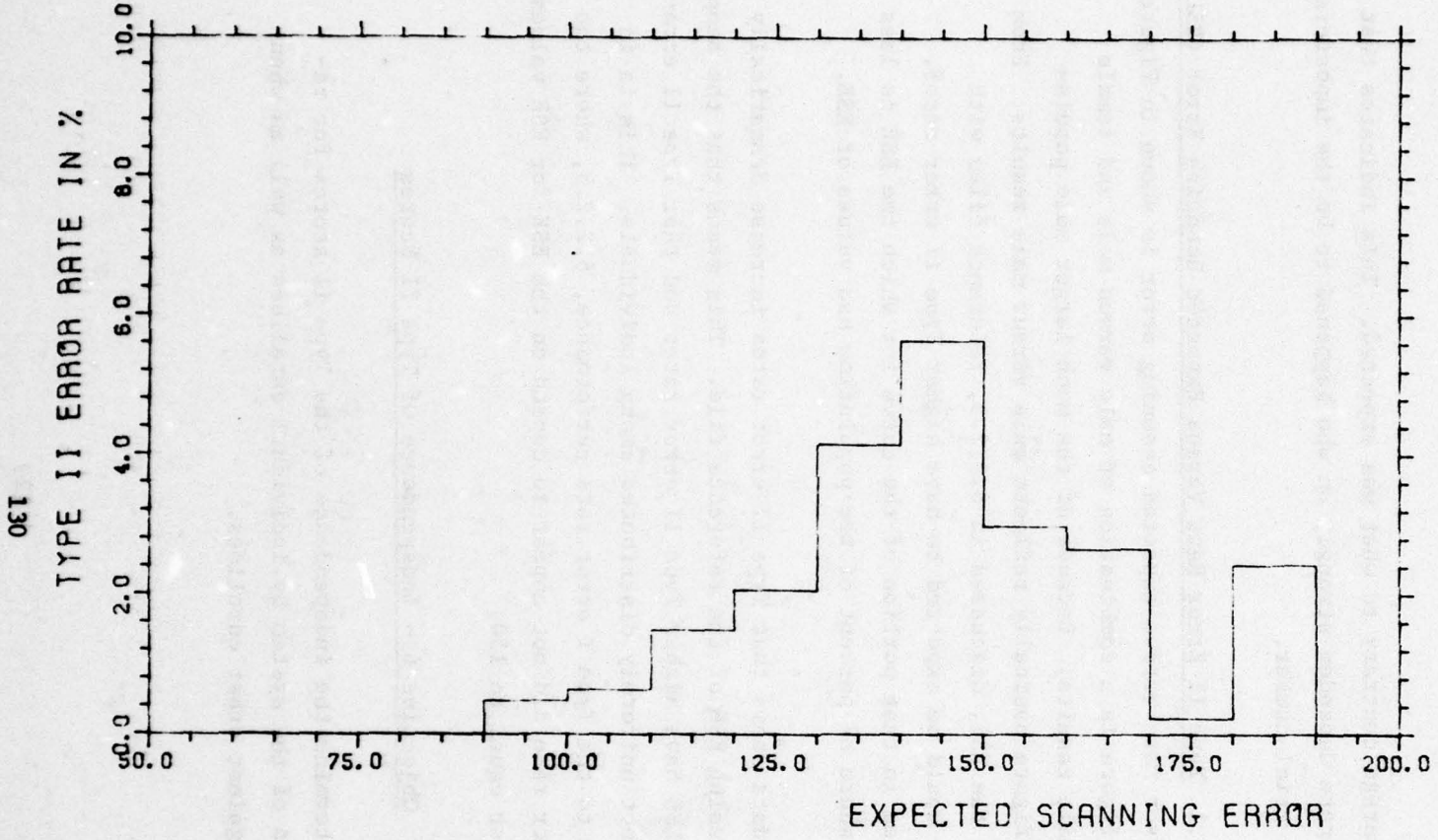


FIGURE 24 TYPE II ERROR RATE VERSUS EXPECTED SCANNING ERROR

6.2.6.1 Independence Of Type II Errors Versus Individuals. Table LIII shows the number of people who had greater than 5, 10, and 15 percent Type II error rates. The number of errors these people had is tabulated. These numbers are indicated, also, as a percent of the total population enrolled and as a percent of the total errors, respectively. It can be seen from the table that a small percent of the population accounted for a disproportionately large percent of the errors. Clearly, the distribution of errors among users is not uniform. This agrees with the Type II versus ESE results (6.2.5.3).

TABLE LIII

PERCENT OF TOTAL POPULATION AND TOTAL ERRORS AS FUNCTIONS OF THE TYPE II ERROR RATE

	Type II Error Rate		
	5%	10%	15%
No. of People	24	12	4
% of Total Population Enrolled (274)	8.8%	4.4%	1.5%
No. of Errors	600	391	149
% of Total Errors (2431)	24.7%	16.1%	6.1%

6.2.6.2 Distribution Of Type II Decision Function Scores. A distribution of frequency of occurrence versus decision function score, for all users in PE and Normal, can be plotted for intruders who registered (see 4.2.6.1 for a discussion of registering) against a reference file. The cumulative decision function was used. To compare the distribution for succeeding phrases, each distribution is divided by the total number of times that the phrase registered. The normalized distributions of frequency of occurrence versus decision function score are plotted on top of the other for each phrase in Figure 25.

The means of the distributions are 182.2, 183.6, 182.2, 178.8 for phrase 1 through 4, respectively. The standard deviations of the distributions are 24.8, 20.7, 20.0, 19.6 for phrases 1 through 4, respectively. Figure 25 was derived from the data with limiting at 200. For the data without limiting, the means would either stay the same or increase as shown in Figures 26-29 where more of the data lies above 200 for each succeeding phrase. Increasing means with increasing phrase number is expected. The occurrences shown here are not the total number of intruder attempts, but only those attempts in which the intruders speech pattern registered against the reference pattern.

6.2.7 Objective 7 - Sensitivity Analysis Of Type I And Type II Errors to Thresholds

To determine the sensitivity of Type I and Type II Error variations to changing thresholds.

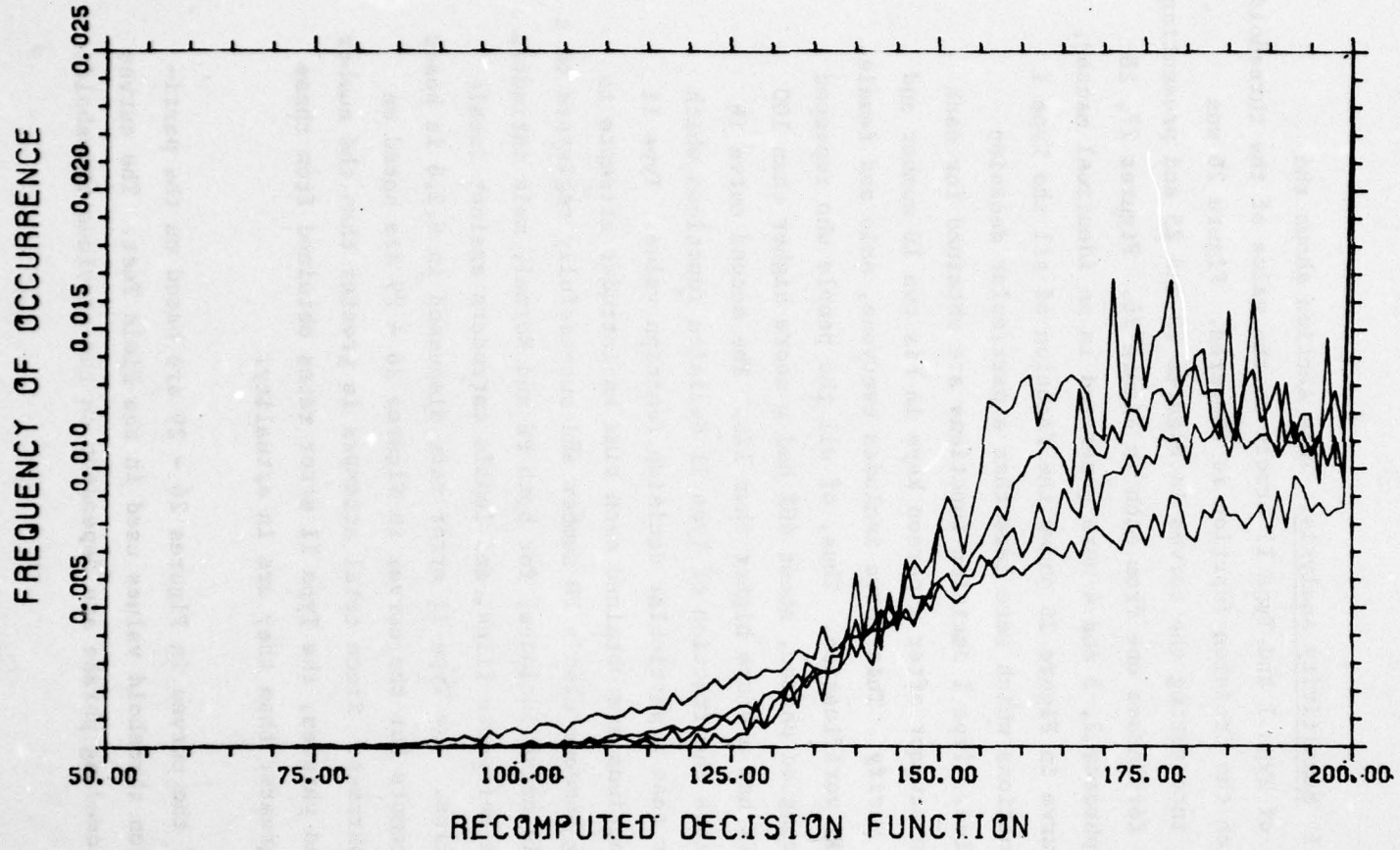


FIGURE 25 FREQUENCY OF OCCURRENCE VS. RECOMPUTED DECISION FUNCTION TYPE II

6.2.7.1 Sensitivity Analysis. This section shows the sensitivity of Type I and Type II errors to the value of the threshold against which the decision function is compared. Figure 26 was obtained by integrating the curves in Figures 22 and 25 and presenting the results for phrase one from each in Figure 26. Figures 27, 28 and 29 for phrases 2, 3 and 4 were generated in an identical manner. The first curve in Figure 26 shows the fraction of all the Type I decision functions which were less than a particular decision function value. Type I decision functions are obtained for each verification attempt after a person keys in his own ID number and attempts to verify. This curve includes everyone, male and female, PE and Normal verification. Thus, of all the people who repeated the first prompted phrase, about 40% had a score higher than 100 and about 20% had a score higher than 120. The second curve in Figure 26 shows the fraction of Type II decision functions which were greater than a particular decision function value. Type II decision functions are obtained each time an intruder attempts to verify using someone else's ID number and successfully registers on a phrase. This curve includes, for both PE and Normal, male intruders against male reference files, and female intruders against female reference files. The Type II error rate discussed in 6.2.6 is based on total attempts but the curves in Figures 26 - 29 are based on registered phrases. Since total attempts is greater than the number of registered phrases, the Type II error rates obtained from these curves are greater than they are in actuality.

Each of the curves in Figures 26 - 29 are based on the particular decision threshold values used in the Field Test. The curves for each succeeding phrase are dependent on the previous thresholds

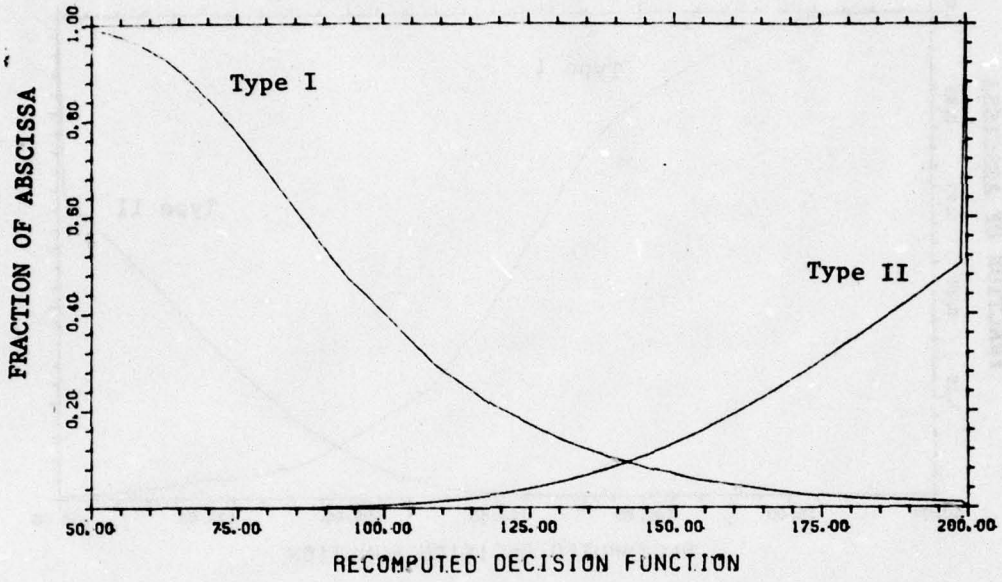


FIGURE 26 FRACTION OF TYPE I "RECOMPUTED DECISION FUNCTION SCORE" (RCDFS) GT AND TYPE II RCDFS LT ABSCISSA PHRASE 1

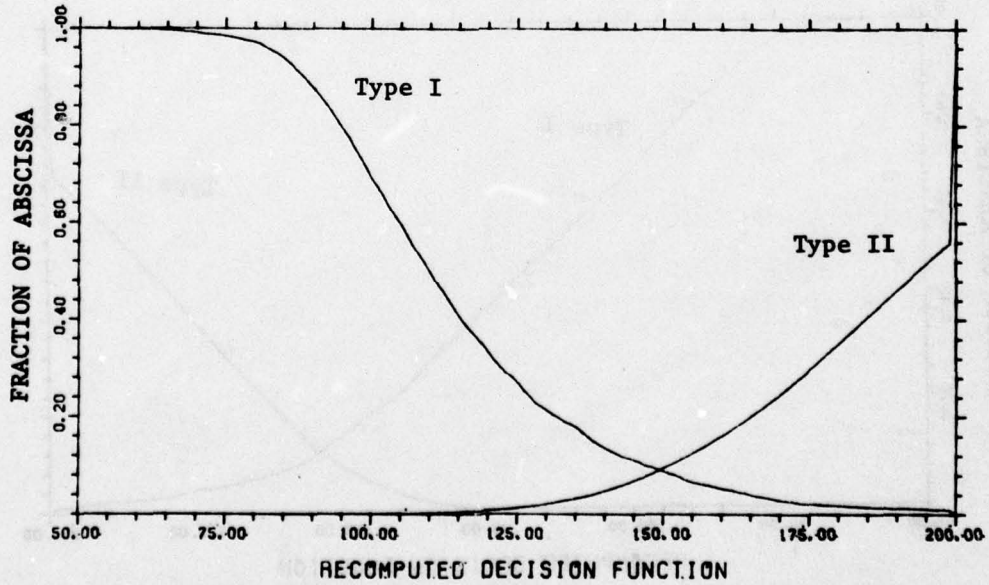


FIGURE 27 FRACTION OF TYPE I RCDFS GT AND TYPE II RCDFS LT ABSCISSA PHRASE 2

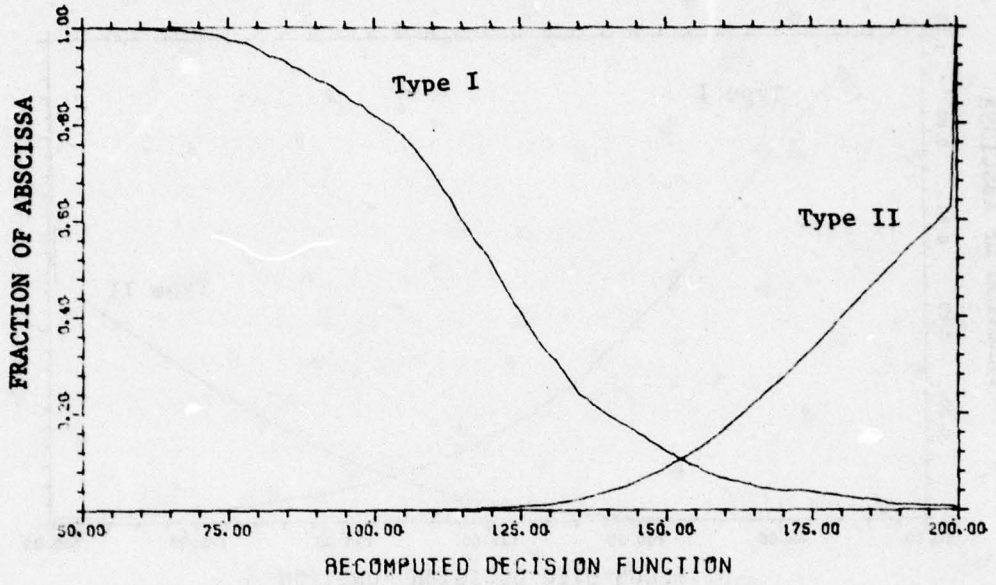


FIGURE 28 FRACTION OF TYPE I RCDFS GT AND TYPE II RCDFS LT ABSCISSA PHRASE 3

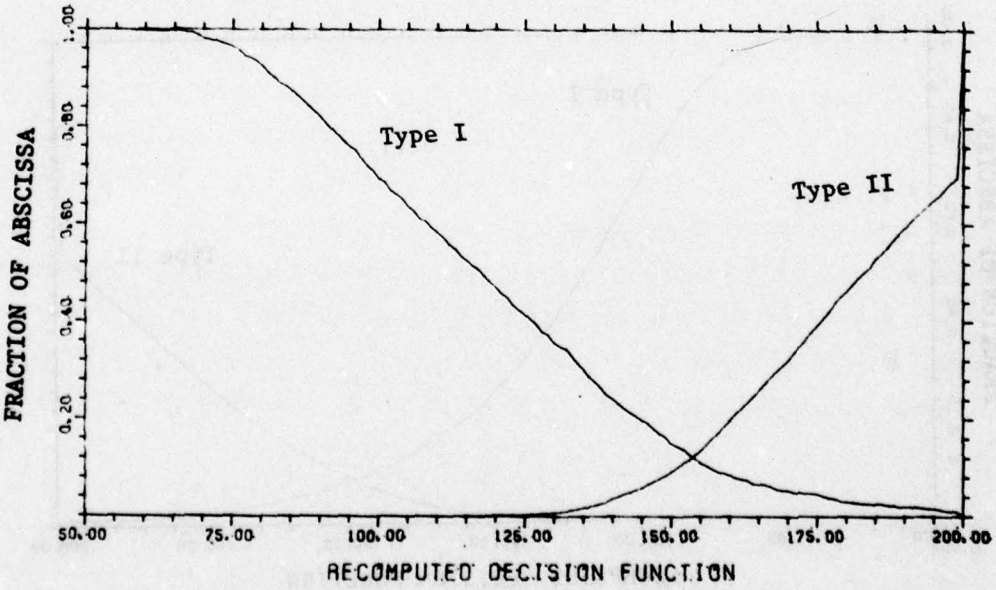


FIGURE 29 FRACTION OF TYPE I RCDFS GT AND TYPE II RCDFS LT ABSCISSA PHRASE 4

as that threshold determines the available population that attempts the next phrase. Therefore, these curves can only be used to predict the performance of the system for small threshold changes. A discussion similar to the one used in 4.2.7.1 can be used to get an initial estimate of the new thresholds needed to get the desired error rates. The actual error rates can then be determined using the recorded data. This has not been done. The use of misregistered phrases as a discriminant must also be taken into account.

6.2.8 Objective 8 - Verification Time Analysis

To determine the average time required for verification and the variance about this time.

6.2.8.1 Service Time (Verification Time). Service time is made up of keyboard time, verification time, first door opening-closing time, second door opening-closing time, and dead time. These last three times will not be discussed here. The average time to keystroke three digits at the keyboard and hit the SEND key was 1.81 seconds. The standard deviation about this time was 0.79 second. The Field Test used only three-digit ID numbers as opposed to four-digit ID numbers in Phases I and II. Also, the keyed numbers were not vocalized by the computer during the field test. For these two reasons this time is smaller than Phases I and II keystroke time.

Table LIV shows a breakdown of the number of decisions, including those that did not verify, by phrases number. From Table LIV it was determined that an average of 5.22 phrases were required to verify

during PE while an average of only 1.62 phrases were required to verify in Normal verification. The average response time per phrase was 1.928 seconds. The standard deviation about this time was 0.40 second. Thus, the average verification time was computed (as described in 4.2.8.1) to be 19.98 seconds during PE but only 6.2 seconds during Normal verification. These times include the 1.9 seconds required to prompts a phrase.

Thus, after an individual has four successful verification attempts and is in Normal verification, the average time for keystroking and verification is 8.0 seconds. In an operational system only a small number of users would be new to the installation and be in PE at any one time. A discussion of throughput is presented in Volume V of this report.

TABLE LIV

NUMBER OF DECISIONS VERSUS PHRASE NUMBER

Phrase Number	1	2	3	4	5	6	7	8	9	10	11	12
Post Enrollment	-	-	-	495	73	31	4	45	32	31	23	9
Normal	3985	1633	532	154	21	15	25	50	20	2	0	0

6.3 FIELD TEST CONCLUSIONS

The ASV algorithm used in the field test was the same as the Phase I with the addition of recycling in the normal mode of operation. As discussed in the report, recycling decreased the Type I error rate slightly but increased the Type II error rate slightly.

The results not using recycling indicate a Type I error rate less than the maximum acceptable rate of 1% but a Type II error rate greater than the maximum acceptable rate of 2%.

The Type II error rates for all users in Normal for Phase I and Field Tests were 1.03% and 3.27%, respectively. Since this result was unexpected, it is discussed further. Figure 30 indicates the Type II error rate versus trial where the error rates were calculated for groups of 5 trials. Phase I Type II error rates are also shown. The 90% confidence intervals (using chi-squared) are indicated on both curves. As seen from the curves, the confidence intervals for the two tests do not overlap indicating that the error rates were obtained from different populations.

Figure 31 shows the Normal verification cumulative Type II error rate versus speaker average. Since it is desired that the Type II error rate be kept under 2%, it can be seen from the curve that it is necessary, in this case, to re-enroll people whose speaker average becomes greater than 135. This was not done in the test. A re-enrollment criteria exists for individual Type I error rates. Therefore, the establishment of a re-enrollment criteria for individual Type II error rates is necessary to come close to meeting the desired Type II error rate.

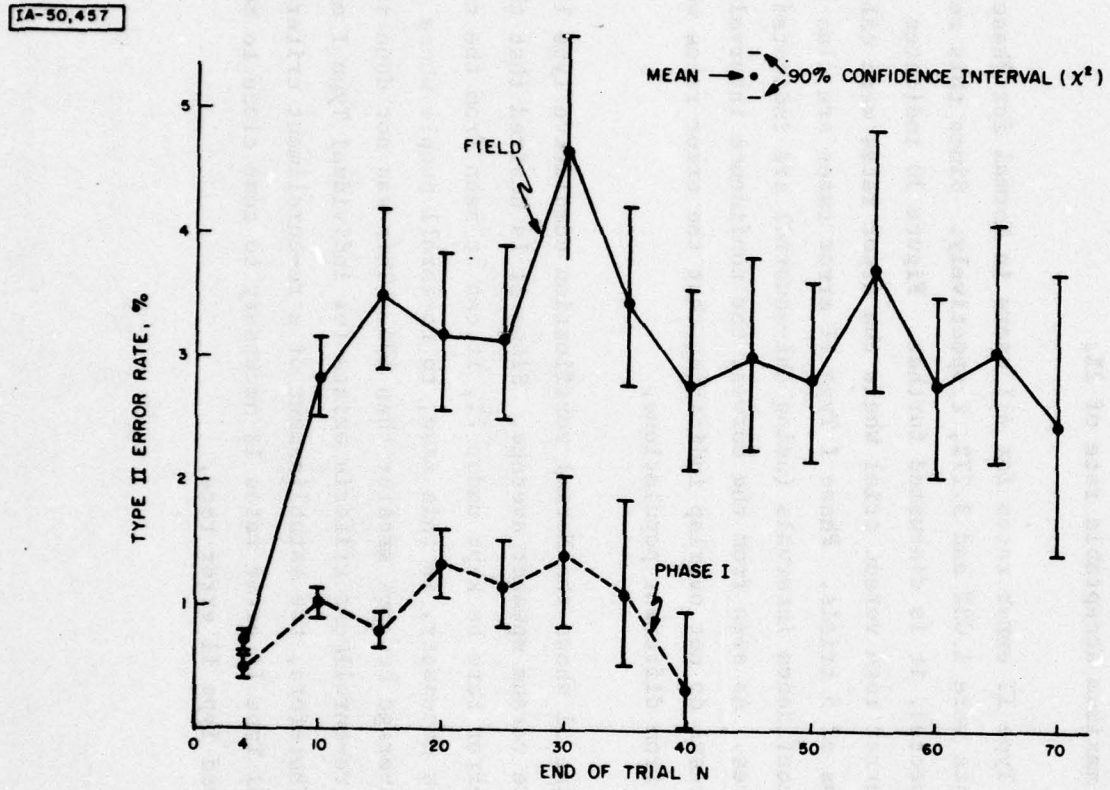


Figure 30 PHASE I AND FIELD TYPE II ERROR RATE VERSUS TRIAL

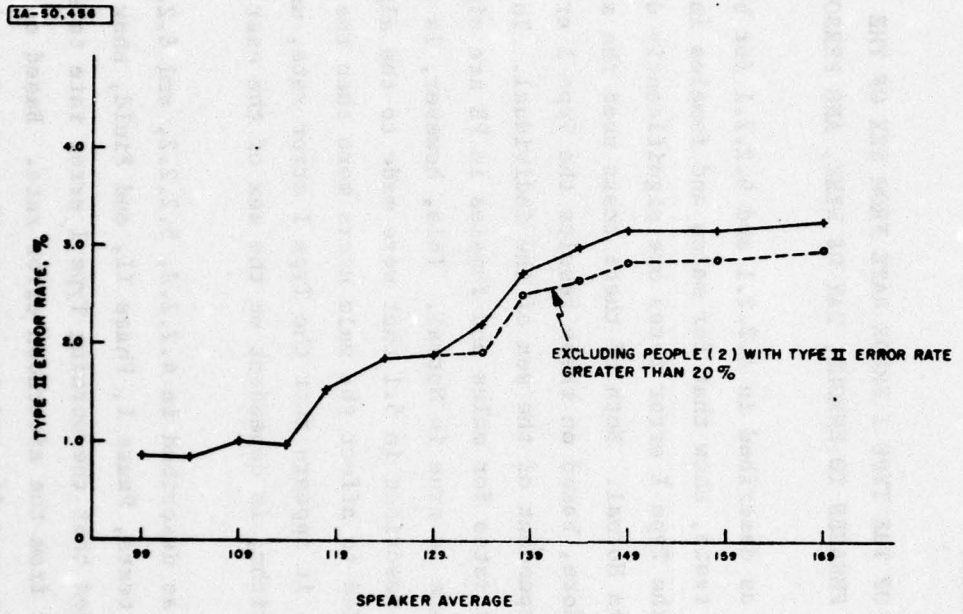


Figure 31 FIELD TEST NORMAL VERIFICATION CUMULATIVE TYPE II ERROR RATE VS SPEAKER AVERAGE

7.0 COMPARATIVE RESULTS

The results of the Phase I, Phase II, and Field tests are compared in this section.

7.1 INDEPENDENCE OF THE TYPE I ERROR RATE FROM SEX OF THE INDIVIDUAL, TIME OF DAY, PHRASES TO ENROLL, DAY OF WEEK, AND PERSONAL STATISTICS

The results, as described in 4.2.2.1 and 6.2.2.1 for both the Phase I and Field tests, show that for males and females in PE it cannot be said that the Type I error rates are significantly different. The same is true in Normal. Both of these tests used the same ASV algorithm. Therefore, based on these results the Type I error rate appears to be independent of the sex of the individual. In Phase II, however, the error rates for males and females in PE are significantly different. The same is true in Normal. This, however, is attributed to the changes as described in 5.1 that were made to the algorithm. These changes tended to affect the male users more then the female users. Therefore, it appears that the Type I error rate, using the Phase II ASV algorithm, is dependent on the sex of the user.

The results, as described in 4.2.2.2, 5.2.2.2, and 6.2.2.2, for each of the three tests, Phase I, Phase II, and Field, show that it cannot be maintained that the morning Type I error rate is significantly different from the afternoon error rate. Based on the agreement of the results as indicated it appears that the Type I error rate is independent of the time of day.

In Phase I, as described in 4.2.2.6, phrases required to enroll were combined into two groups. It was determined that it could not be maintained that the error rate for these groups are significantly different. In the Field Test as described in 6.2.2.6, it was determined that the Type I error rates of the two groups are significantly different. There was insufficient data in Phase II to do an analysis in this category. Based on the limited data and the dichotomy of the results it appears that additional testing is required to determine the dependence, if any, of Type I error rate on phrases required to enroll.

The results, as described in 4.2.2.7, 5.2.2.7, and 6.2.2.7, indicate, respectively, that the Type I error rate as a function of the day of the week does not have any apparent trend, tends to increase as the week progresses, and tends to decrease as the week progresses. Since sufficient data seems to be available, the apparent random nature of the results suggest that the Type I error rate is independent of the day of the week.

Type I errors versus personal statistics were examined for the test population of Phase II and the Field Tests and there appeared to be no significant effect on the ASV performance.

7.2 PHASE I VERSUS PHASE II

Phase I performance in the Normal mode was quite acceptable. However, for PE, the Type I error rate was higher than need be and

the Type II error rate was lower than need be. Based on these results, changes, such as adjusting the threshold, were made to the algorithm to correct these deficiencies.

As seen from the Phase II results (summarized in Table I) the changes in the algorithm shifted the performance too far. Not only were the Type I and Type II error rates shifted too far in PE, but so were the error rates in Normal operation. Some of the drop in the PE Type I error rate was due to the fact that the same population participated in both phases. Almost all of the shift in the Type II error rate occurred for male versus male. This was because the changes made to the algorithm affected males almost exclusively.

The expected scanning error was about 20 points higher for females than it was for males in both Phase I and Phase II. Change number 2, see 5.1, therefore, affected the male population almost exclusively. This is manifested by comparing the Type II error rates for males and females in Phase I and Phase II where the male rate increased by a factor of four while the female rate did not change significantly. This in combination with change 3, especially in PE, improved the Type I error rate, but at the expense of the Type II error rate.

The advantage of recycling phrases in Normal as described in 5.2.5.1 is not clear, especially when considering Type II errors. As a matter of fact, it may be costly. Even in PE, it is not clear that there is an advantage.

7.3 PHASE I VERSUS FIELD TEST

Since little time existed between the end of Phase II and the start of Field testing, it was decided to use the Phase I algorithm in the Field Test. The Field Test results, summarized in Table I, are, therefore, compared with the Phase I results. The Type I performance, in Normal operation, compares favorably with that observed during Phase I. The PE results show that the users in the field had more difficulty using the system initially than did the Phase I users, but that they learned quickly.

The rise in the Type II error rate seen in Table I was unexpected. Part of the rise (0.5% to 0.75%) can be attributed to a higher ESE average in the Field Test (123) than in Phase I (114). As noted earlier, there is a correlation between ESE and the Type II error rate. Also, fewer users in the Field Test reached the Normal mode of operation so that the influence of a few files with high Type II error rates (ten users had a Type II error rate greater than 10%) may have had a significant effect. For example, one user with an ESE of 117 had a Type II error rate of 10.61% while three other users with the same ESE had rates of 2.29%, 3.05% and 3.03%. In another case, for ESE's of 129 and 130, one user had a rate of 21.15% while three others had rates of 0%, 1.15%, and 0%.

Since the Field Test sample population appears to differ significantly from the Phase I sample population, a new procedure may have to be added to the system. That is, to re-enroll users who have an ESE greater than 140 at the end of PE. Experience has shown that re-enrollment drops the ESE, sometimes dramatically. Analysis of the Field Test data has shown that eliminating all test subjects with an ESE above 140 reduces the Type II error rate from 3.26% to 2.66%. If

the ESE threshold is set at 135, the Type II error rate would meet the required 2%. The test showed that less than 5% of the users would have to be re-enrolled because of high Type I errors, and another 5% to 10% would have to be re-enrolled due to high ESE's, i.e., Type II errors.

With re-enrollment of the high ESE population, the Type II error rate of a large mixed field population should be less than 2.75% for the period following PE. Experience at Texas Instruments shows that the ESE continues to go down, even after 2,000 trials (about two years) so that the Type II error rate should also decrease.

7.4 OTHER TYPE II TESTING

Limited, but more sophisticated Type II testing, was conducted against the ASV system using MITRE personnel, college drama and speech students and faculty. The results are reported in a separate report. Since the ASV system uses a fixed common vocabulary, the random Type II testing is a meaningful test of the system. The dedicated mimic will have to learn to change his speech to match the formant frequency structure of the person he is trying to mimic. Earlier tests conducted by Bell Laboratories and Texas Instruments showed that professional mimics could double the random Type II error rate. Although the use of recorded data from an enrolled user is a viable option, it would take elaborate equipment. This is because the phrases are prompted in a random sequence, the user has a limited time (4 sec) in which to respond, and the spectrum of the tape recorder/player must match the original speaker. A one-time series of tests to determine the success of record voice data should be conducted.

APPENDIX A

DETERMINING IF THE ERROR RATES FROM TWO GROUPS ARE SIGNIFICANTLY DIFFERENT

In BISS data analysis it is often desired to determine if the error rates from two groups are significantly different. The BISS data provides only samples from different groups. The F test provides a test to determine statistically if the underlying populations of the different groups do not have the same error rates.

The binominal distribution

$$P_N(k) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k} \qquad (1)$$

is the probability of having k errors in N trials where the probability of error is p and that of no error is 1 - p. Each trial is taken to be independent. In the limiting case where

$$p \quad \text{is small } (p \ll 1)$$

$$N \quad \text{is large } (N \gg 1) \text{ and}$$

$$N \gg k, \qquad (2)$$

the binomial distribution may be approximated by the Poisson distribution as [4]

$$\frac{N!}{k!(n-k)!}\, p^k\, (1-p)^{N-k} \approx \frac{(Np)^k\, e^{-Np}}{k!} \tag{3}$$

From the relationship

$$\Gamma(x+1) = x!$$

equation (3) may be rewritten as

$$P_k(Np) = \frac{(Np)^{(k+1)-1}\, e^{-Np}}{\Gamma(k+1)} \tag{4}$$

The continuous variable Np is a Gamma variate with parameter $k+1$ and its distribution, equation (4), is a Gamma distribution. [5]

The probability that the variable Np is in the interval dNp is

$$dP = \frac{(Np)^{(k+1)-1}\, e^{-Np}}{(k+1)}\, dNp \tag{5}$$

Substituting

$$k+1 = \frac{n}{2} \tag{6}$$

into equation (5) yields:

$$dP = \frac{2^{\frac{1}{2}(n-2)}\,(Np)^{\frac{1}{2}(n-2)}\,e^{-Np}}{2^{\frac{1}{2}(n-2)}\,\Gamma(n/2)}\; d(2Np/2)$$

$$= \frac{(2Np)^{\frac{1}{2}(n-2)}\,e^{-Np}}{2^{n/2}\,\Gamma(n/2)}\; d(2Np) \qquad (7)$$

The probability density function

$$P_C(2Np) = \frac{(2Np)^{\frac{1}{2}(n-2)}\,e^{-Np}}{2^{n/2}\,\Gamma(n/2)}$$

is the chi-square distribution for the variate 2Np with n degrees of freedom.

Consider that $2N_1p_1$ and $2N_2p_2$ are independent, and that each is chi-square distributed with n_1 and n_2 degrees of freedom, respectively. Then

$$T(n_1,n_2) = \frac{2N_1p_1 \;/\; n_1}{2N_2p_2 \;/\; n_2} \qquad (8)$$

is the F variate and its distribution is the F distribution.

Replacing n with $2(k + 1)$ from equation (6), equation (8) becomes

$$T[2(k_1 + 1),\ 2(k_2 + 1)] = \frac{2N_1P_1(2k_2 + 2)}{2N_2P_2(2k_1 + 2)} \tag{9}$$

$$= \frac{N_1P_1(k_2 + 1)}{N_2P_2(k_1 + 1)}$$

if the error rates p_1 and p_2 of the underlying populations are hypothesized to be equal, then equation (9) becomes

$$T(2k_1 + 2,\ 2k_2 + 2) = \frac{N_1(k_2 + 1)}{N_2(k_1 + 1)}$$

This hypothesis can be tested by seeing how often it will happen that the F variate with parameters $2k_1 + 2$ and $2k_2 + 2$, $F(2k_1 + 2, 2k_2 + 2)$, will take on values $\geq (N_1k_2 + N_1)/(N_2k_1 + N_2)$.

There are tables[8] of the cumulative F distribution tabulating against a_1, a_2 the values of $F(a_1,a_2)$ such that

$$P = \int_0^{F_P} f[F(a_1,a_2)]\ dF = \int_0^{F_P} f[F(2k_1 + 2,\ 2k_2 + 2)]\ dF \tag{10}$$

where f(F) is the probability density function. From equation (10) it is meant that 100P % of the time the values of $F(2k_1 + 2, 2k_2 + 2)$ will be $\leq F_P$. Thus, if the observed T is less than or equal to F_P, then there is a 100P % "level of confidence" and $T = (N_1k_2 + N_1)/(N_2k_1 + N_2)$ is an F variate and therefore, the underlying hypothesis of equal error rates cannot be rejected.

If P is chosen large enough, then in the case of $T > F_p$ it can be concluded that it is very unlikely that T is F distributed. The hypothesis of equal error rates can then be rejected. The inverse

reasoning is not possible. If $T \leq F_p$, one cannot say that T is F distributed and the hypothesis is true, but only that the hypothesis is not inconsistent with the results.

For example, if N_i are the entry attempts of group i and k_i are the errors in group i, then for

$$N_1 = 1646 \qquad k_1 = 23$$

$$N_2 = 1027 \qquad k_2 = 19$$

$$T[2(23 + 1),\ 2(19 + 1)] = \frac{1646\ (19 + 1)}{1027\ (23 + 1)}$$

or

$$T(48,\ 40) = 1.336$$

From the F_p tables for P = .9, the value of F_p such that

$$.9 = \int_0^{F_P} f[F(48,\ 40)]\ dF$$

is

$$F_p = 1.49$$

For this example, $T \leq F_p = .9$ and therefore, at the 90% confidence level the hypothesis that the two groups have equal error rates cannot be rejected.

In this manner, the F test can be used to determine if the error rates of two different groups are significantly different in a statistical sense.

APPENDIX B

COMPUTING AN UPPER BOUND ON ERRORS

The basis for the formula to compute error bounds in identity verfication is described below. Assume that the number of verification errors for a given number of (trials) entry attempts is statistical equivalent to the number of failures of a piece of equipment for a given number of operating hours. Specifically, failures are replaced with errors and operation hours are replaced with entry attempts. Now, the upper bound on the number of errors at a given confidence level, can be estimated for a fixed number of trials in the same way that the number of failures, at a given confidence level, can be estimated for a fixed number of operating hours.

The formula used to estimate the lower bound of the mean time between failures (MTBF) for a fixed time test is given by[6]

$$MTBF_c = 2T \Big/ [\chi^2_{1-c} (2f + 2)]$$

where

T is time period
c is confidence level, e.g., 90%
f is number of failures
2f + 2 number of degrees of freedom

The upper bound on failure rate is the inverse of the lower bound on MTBF. Thus

$$\text{Upper Bound Failure Rate} = 1/\text{MTBF}_{\text{lower}} = \chi^2\ (2f + 2)/2T$$

The upper bound on the number of failures divided by the time period T is the upper bound on the failure rate

$$\text{Failures}_c = \chi^2_{1-c}\ (2f + 2)/2.$$

Substituting errors, e for failures, f yields

$$\text{Errors}_c = \left[\chi^2_{1-c}\ (2e + 2)\right]/2.$$

An example of the use of this equation follows. On the first entry attempt using the speaker verification system there were 8 failures to verify an authorized individual out of 193 entry attempts leading to an e = 8. The upper bound on the error at 90% confidence is given by (see Reference 8 for chi-square distribution tables).

$$\text{Errors}_{.9} = \chi^2(18)/2 = 26/2 = 13.$$

We can say with 90% confidence that no more than 13 errors out of 193 trials will occur on the first trial.

The best estimate of the error rate is 4.14% and a 90% confidence level on the upper bound of the error rate is 4.14% × 13/8 = 6.73%.

APPENDIX C

DISCUSSION OF TYPE II ERRORS WITH RECYCLING

After publication of this report, an error was found in the analysis program which computed the Type II errors in Post Enrollment with recycling. This appendix describes a technique for estimating the values which have been assigned as "unavailable" without reanalyzing the Type II error data collected during the tests.

Section 5.2.5.1 of this report provides a discussion of the effects of recycling misregistered phrases on Type II errors. The primary reason for recycling is to reduce Type I errors. However, a reduction of Type I error rate results in an increase in the Type II error rate. The sequential decision strategy of the ASV system is designed so that the Type I and Type II error rates are approximately constant for each individual phrase. When recycling is applied, the resulting system Type I error rate should decrease exponentially (Type I error rates for individual phrases multiply), and the system Type II error rate should increase linearly (Type II error rates add) with each successive phrase.

Accordingly, when recycling is applied to the Normal mode strategy (i.e., five phrases prompted instead of four) there should be an attendant 25% increase in the Type II error rate. Analysis of the Normal mode test data shows an increase of 23% in the Type II error rate during Phase II (see Normal mode in Table XXXI or Table XXXI-C; 4779/3874 = 1.23), and an increase of 30% during the Field Test (see Normal mode data in Table LI or Table LI-C; 1124/864 = 1.30); the average increase is 27%.

In the Post Enrollment decision strategy, up to two phrases may be recycled when recycling is applied. An estimate of the upper bound on the Type II error rate with recycling during Post enrollment may be obtained from Post Enrollment without recycling. This may be accomplished by assuming an average increase in Type II error rate of 27% per additional phrase and applying the resulting factor to Post Enrollment data rather than reanalyzing all the real time data gathered during the test. Estimates for Type II errors and error rates for PE with recycling appear in the tables below. (Since up to two phrases are allowed in Post Enrollment, a factor of 1.54 is used to compute recycling values from no recycling data. For example, in Table XXXIII-C, the estimated number of errors with recycling during Post Enrollment for male versus male, 1865, is equal to the number of errors recorded during the test, 1211, times 1.54.)

The approach to recycling described in Section 5.2.5.1 results in an upper bound on the Type II error rate because the probability of a phrase registering given that it has already misregistered is lower than the probability of a new phrase registering (due to the fact that a new phrase is made up of from one to four words which had previously registered).

For the Field Test, in the Normal mode, there were at least 23 instances in which the users required recycling in real time (not all instances were identified in the original processing). Of these, nine verified. There were 40 Type I errors in the Normal mode without recycling as shown in Table XXXIX. With recycling, nine of the 40 cases verified thereby reducing the number of Type I errors to 31 for an improvement of 22.5%. Data for a similar comparison in Phases I and II were not available. Based on this limited comparison of the Field Test data (a drop of 22.5% in the Type I error rate versus a 30% increase in the Type II error), it is not clear that recycling in the Normal mode should be used.

TABLE XXXI-C

TYPE II ERROR RATES-ALL USERS

	Errors		Attempts	Error Rate (%)	
	R*	NR**	R and NR	R	NR
Post Enrollment	1879	1220	19526	9.6	6.25
Normal	4779	3874	90410	5.29	4.28
Total	6658	5094	109936	6.1	4.63

TABLE XXXIII-C

TYPE II ERROR RATES WITH RECYCLING

Category	Errors	Attempts	Error Rate (%)
Post Enrollment			
Male Vs. Male	1865	18846	9.9
Female Vs. Female	14	680	2.1
Normal			
Male Vs. Male	4666	86371	5.40
Female Vs. Female	113	4039	2.80

TABLE LI-C

TYPE II ERROR RATES ALL USERS

	Errors		Attempts	Error Rate (%)	
	R*	NR*	R and NR	R	NR
Post Enrollment	403	262	36904	1.1	0.71
Normal	1124	864	26411	4.26	3.27
Total	1527	1126	63315	2.4	1.78

*R results using recycling.
NR results not using recycling.

REFERENCES

1. Doddington, G.R., "Speaker Verification," UI-713804-F, Texas Instruments Inc., 13,500 North Central Expressway, Dallas, Texas 75222, July 1974.

2. Doddington, G.R., "Speaker Verification II," UI-713804-F, Texas Instruments Inc., 13,500 North Central Expressway, Dallas, Texas 75222, 15 September 1975.

3. Rosenberg, A.E., "Automatic Speaker Verification - A Review," Procedure of IEEE, Vol. 64, No. 4, April 1976, Pages 480-482.

4. Meyer, Stuart L., "Data Analysis for Scientists and Engineers," John Wiley and Sons, Inc. 1975, pp. 202-212, 285-286.

5. Weatherburn, G.E., "A First Course in Mathematical Statistics," Cambridge Press, 1949, Chapters VIII and IX.

6. Epstein, B., "Estimation From Life Test Data," IRE Transactions on Reliability and Quality Control, April 1960.

7. Brandt, S., "Statistical and Computational Methods in Data Analysis," North-Holland Publishing Co., 1970, pp. 210-213.

8. Handbook of Tables for Probability and Statistics, Beyer, W.H., editor, The Chemical Rubber Co., 2,310 Superior Avenue, Cleveland, Ohio 44114, pp. 234 and 241.